A One Covariate at a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models*

A. Chudik
Federal Reserve Bank of Dallas

G. Kapetanios
King's College, London

M. Hashem Pesaran
Department of Economics & USC Dornsife INET, University of Southern California, USA
and Trinity College, Cambridge, UK

February 23, 2018

Abstract

This paper provides an alternative approach to penalised regression for model selection in the context of high dimensional linear regressions where the number of covariates is large, often much larger than the number of available observations. We consider the statistical significance of individual covariates one at a time, whilst taking full account of the multiple testing nature of the inferential problem involved. We refer to the proposed method as One Covariate at a Time Multiple Testing (OCMT) procedure, and use ideas from the multiple testing literature to control the probability of selecting the approximating model, the false positive rate and the false discovery rate. OCMT is easy to interpret, relates to classical statistical analysis, is valid under general assumptions, is faster to compute, and performs well in small samples. The usefulness of OCMT is also illustrated by an empirical application to forecasting U.S. output growth and inflation.

Keywords: One covariate at a time, multiple testing, model selection, high dimensionality, penalised regressions, boosting, Monte Carlo experiments
JEL Classifications: C52, C55

*We are grateful to the Editor (Elie Tamer) and three anonymous referees for constructive comments on earlier versions of the paper. We have also benefited from helpful comments by Jinchi Lv, Yingying Fan, Essie Maasoumi, Mahrad Sharifvaghefi, Ron Smith, and seminar participants at the Bank of England, Federal Reserve Bank of Dallas, Third IAAE Conference at the University of Milano-Bicocca, Michigan State University, University of Michigan (Department of Statistics), University of California, Irvine, University of Kent, University of Messina, University of Southern California, and University of Warwick. The views expressed in this paper are those of the authors and do not necessarily represent those of the Federal Reserve Bank of Dallas or the Federal Reserve System.

1 Introduction

This paper contributes to the literature by proposing a multiple testing procedure to model selection in high dimensional regression settings. The goal of the proposed procedure is to select an approximating model that encompasses the true model, and does not contain any noise variables that are uncorrelated with signal (true) variables. We use ideas from the multiple testing literature to control the probability of selecting the approximating model, the false positive rate and the false discovery rate. We refer to the proposed method as One Covariate at a Time Multiple Testing (OCMT) procedure. OCMT is computationally simple and fast even for extremely large data sets.

Our approach is to be contrasted to penalised regressions where the vector of regression coefficients, $\boldsymbol{\beta}$, of a regression of y_t on $\boldsymbol{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})'$, known as the active set, is estimated by $\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}} = \text{argmin}_{\beta}[\sum_{t=1}^{T}(y_t - \boldsymbol{x}'_{nt}\boldsymbol{\beta})^2 + P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})]$. $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ is a penalty function that penalises $\boldsymbol{\beta}$, while $\boldsymbol{\lambda}$ is a vector of tuning parameters to be set by the researcher. A variety of penalty functions have been considered, yielding a wide range of penalised regression methods. Chief among them is Lasso, where $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ is chosen to be proportional to the L_1 norm of $\boldsymbol{\beta}$. This has subsequently been generalised to penalty functions involving L_q, $0 \leq q \leq 2$, norms. While these techniques have found considerable use in econometrics,[1] their theoretical properties have been mainly analysed in the statistical literature starting with the seminal work of Tibshirani (1996) and followed up with important contributions by Fan and Li (2001), Antoniadis and Fan (2001), Efron et al. (2004), Zhou and Hastie (2005), Candes and Tao (2007), Lv and Fan (2009), Bickel et al. (2009), Zhang (2010), Fan and Lv (2013) and Fan and Tang (2013). Despite considerable advances made in the theory and practice of penalised regression, there are still a number of open questions. These include the choice of the penalty function and tuning parameters. A number of contributions, notably by Fan and Li (2001) and Zhang (2010), have considered the use of nonconvex penalty functions with some success.[2]

Like penalised regressions, OCMT is valid when the underlying regression model is sparse. Further, it does not require the $\boldsymbol{x}_{nt}$ to have a sparse covariance matrix, and is applicable even if the covariance matrix of the noise variables, to be defined below, is not sparse. Of course, since OCMT is a model selection device, well known impossibility results for the uniform validity of post-selection estimators, such as those obtained in Fan and Pötscher (2006) and Fan and Pötscher (2008), apply. The main idea is to test the statistical significance of the net contribution of all n available potential covariates in explaining y_t individually, whilst taking full account of the multiple testing nature of the problem under consideration. All covariates

[1] A general discussion of high-dimensional data and their use in microeconomic analysis can be found in Belloni et al. (2014a).

[2] As an alternative to penalized regression, a number of procedures developed in the machine learning literature such as boosting, regression trees, and step-wise regressions are also widely used. See, for example, Friedman et al. (2000), Friedman (2001), Buhlmann (2006) and Fan and Lv (2008).

with statistically significant net contributions are then selected *jointly* to form an initial model specification for y_t. Unlike boosting and other greedy algorithms, our procedure is not sequential and selects in a single step all covariates whose t-ratios exceed a given threshold. A second stage will be needed only if there exist hidden signals, in the sense that there are covariates whose net contribution to y_t is zero, despite the fact that they belong to the true model for y_t. To allow for the possibility of hidden signals, we propose a multi-stage version, where OCMT is repeated by testing the statistical contribution of the remaining covariates, not selected in the first stage, again one at a time, to the unexplained part of y_t. We will show that this multi-stage process converges in a finite number of steps, since the number of hidden signals cannot rise with n. In a final step all statistically significant covariates, from all stages, are included as joint determinants of y_t in a multiple regression setting. Whilst the initial regressions of our procedure are common to boosting (see Buhlmann (2006)) and to the screening approach discussed in Fan and Lv (2008), Huang et al. (2008), Fan et al. (2009) and Fan and Song (2010), OCMT provides an inferentially motivated stopping rule without resorting to the use of information criteria, or penalised regression after the initial stage.

Related sequential model selection approaches have been proposed, among others, by Fithian et al. (2014), Tibshirani et al. (2014) and Fithian et al. (2015). In the context of linear regression, these methods build regression models by selecting variables from active sets, based on a sequence of tests. The use of multiple testing, implies that the choice of critical values, used at every testing step in the sequence, is crucial and there have been a number of important contributions, in this respect, including Li and Barber (2015) and G'Sell et al. (2016).

We provide theoretical results for the proposed OCMT procedure under relatively mild assumptions. In particular, we do not assume either a fixed design or time series independence for $\boldsymbol{x}_{nt}$ but consider a martingale difference condition for the cross-products $x_{it}x_{jt}$ and $\boldsymbol{x}_{nt}u_t$, where u_t is the error term of the true model. While these martingale difference conditions are our maintained assumption, we also provide theoretical arguments that allow the covariates to follow mixing processes. We establish theoretical results on the true positive rate, the false positive rate, the false discovery rate, and the norms of the coefficient estimate as well as the regression error.

We investigate the small sample properties of the proposed estimator and compare its performance with a number of penalised regressions (including Lasso and Adaptive Lasso), and boosting techniques. We consider data generating processes with and without lagged values of y_t, and carry out a large number of experiments. Although no method uniformly dominates, the results clearly show that OCMT does well across a number of dimensions. In particular, OCMT is very successful at eliminating noise variables, whereas it is still quite powerful at picking up the signals. It is outperformed by Lasso and Adaptive Lasso for a small fraction of experiments only. The relative performance of OCMT is also illustrated in an empirical application to forecasting U.S. output growth and inflation.

The paper is structured as follows: Section 2 explains the basic idea behind the OCMT method and introduces the concepts of the true and approximating models. Section 3 provides a formal description of the OCMT method and derives its asymptotic properties. Sections 4 presents a number of extensions. Section 5 gives the details of the Monte Carlo experiments and the summary of the simulation results. Section 6 presents the empirical application, and Section 7 concludes. Online supplement, organized in three parts, provide additional theoretical results and proofs, a complete set of Monte Carlo results for all the experiments conducted, and additional empirical findings.

Notations: Generic positive finite constants are denoted by C_i for $i = 0, 1, 2, ...$. They can take different values at different instances. If $\{f_n\}_{n=1}^{\infty}$ is any real sequence and $\{g_n\}_{n=1}^{\infty}$ is a sequences of positive real numbers, then $f_n = O(g_n)$, if there exists a positive finite constant C_0 such that $|f_n| / g_n \leq C_0$ for all n. $f_n = o(g_n)$ if $f_n/g_n \to 0$ as $n \to \infty$. If $\{f_n\}_{n=1}^{\infty}$ and $\{g_n\}_{n=1}^{\infty}$ are both positive sequences of real numbers, then $f_n = \ominus(g_n)$ if there exists $N_0 \geq 1$ and positive finite constants C_0 and C_1, such that $\inf_{n \geq N_0}(f_n/g_n) \geq C_0$, and $\sup_{n \geq N_0}(f_n/g_n) \leq C_1$. $\to_p$ denotes convergence in probability as $n, T \to \infty$.

2 True and Approximating Models and OCMT

Consider the data generating process (DGP),

$$y_t = \mathbf{a}' \boldsymbol{z}_t + \textstyle\sum_{i=1}^{k} \beta_i x_{it} + u_t, \tag{1}$$

where $\boldsymbol{z}_t$ is a known vector of pre-selected variables, $x_{1t}, x_{2t}, ..., x_{kt}$ are the k unknown *true* or *signal* variables, $0 < |\beta_i| \leq C < \infty$, for $i = 1, 2, ..., k$, and u_t is an error term. It is assumed that $\boldsymbol{z}_t$ and x_{it}, $i = 1, 2, ..., k$, are uncorrelated with u_t at time t. $\boldsymbol{z}_t$ may include deterministic terms such as a constant, linear trend and dummy variables, and/or stochastic variables, possibly including common factors and lagged values of y_t, that are considered crucial for the modelling of y_t, and are selected based possibly on *a priori* theoretical grounds.

Further suppose that the k signals are contained in a set $\mathcal{S}_{nt} = \{x_{it}, i = 1, 2, ..., n\}$, with n being potentially larger than T, which we refer to as the *active* set.[3] In addition to the k signals, the active set is comprised of *noise* variables that have *zero* correlations with the signals once the effects of $\boldsymbol{z}_t$ are filtered out, and a remaining set of variables that, net of $\boldsymbol{z}_t$, are correlated with the signals. We refer to the latter as *pseudo-signals* or *proxy* variables, since they can be falsely viewed as signals.

[3] We assume that the signal variables are contained in the active set. Nevertheless, OCMT can be applied even if the active set does not contain all of the signal variables. It is clear that in such a setting the true model or a model that contains the true model cannot be identified. However, OCMT will still weed out the variables that are uncorrelated with the signals. In support of this, we provide Monte Carlo evidence in Section 5 of the online MC supplement, based on a Monte Carlo experiment suggested to us by a referee.

The OCMT procedure considers the least squares (LS) regression of y_t on $\boldsymbol{z}_t$ and the regressors in the active set *one at the time*. Let t_i be the t-ratio of x_{it} in the regression of y_t on $\boldsymbol{z}_t$ and x_{it}, for $i = 1, 2, ..., n$,

$$t_i = \frac{T^{-1/2}\boldsymbol{x}_i'\mathbf{M}_z\boldsymbol{y}}{\hat{\sigma}_i\sqrt{T^{-1}\boldsymbol{x}_i'\mathbf{M}_z\boldsymbol{x}_i}} = \frac{T^{-1/2}\boldsymbol{x}_i'\mathbf{M}_z\boldsymbol{\mu}}{\hat{\sigma}_i\sqrt{T^{-1}\boldsymbol{x}_i'\mathbf{M}_z\boldsymbol{x}_i}} + \frac{T^{-1/2}\boldsymbol{x}_i'\mathbf{M}_z\boldsymbol{u}}{\hat{\sigma}_i\sqrt{T^{-1}\boldsymbol{x}_i'\mathbf{M}_z\boldsymbol{x}_i}} = t_{i,\mu} + t_{i,u}, \tag{2}$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iT})'$ and $\boldsymbol{y} = (y_1, y_2, ..., y_T)'$ are $T \times 1$ vectors of observations on x_{it} and y_t, respectively, $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_T)'$, $\mu_t = \sum_{i=1}^{k}\beta_i x_{it}$, $\boldsymbol{u} = (u_1, u_2, ..., u_T)'$, $\mathbf{M}_z = \boldsymbol{I}_T - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_T)'$ is the matrix of observations on $\boldsymbol{z}_t$, and $\hat{\sigma}_i$ is the standard error of the regression of y_t on $\boldsymbol{z}_t$ and x_{it}.

Consider first $t_{i,u}$, defined by (2), which plays a key role in the workings of the OCMT. As $n, T \to \infty$, we rely on $t_{i,u}$ to remain bounded in probability sufficiently sharply so as to allow for multiple testing over very large values of n. We obtain such bounds under a variety of relatively mild assumptions on u_t and x_{it}. For example, we allow u_t to be a martingale difference process and require x_{it} to be uncorrelated with u_t. We do not require x_{it} to be strictly exogenous.

Regarding $t_{i,\mu}$ in (2), we distinguish between the cases where $t_{i,\mu}$ is bounded in probability sufficiently sharply as $n, T \to \infty$ and when it is not. The latter case is of special interest and suggests that x_{it} has power in explaining y_t, net of the pre-selected variables, $\boldsymbol{z}_t$. In such a case, we select x_{it}, and we distinguish between the signal variables, that are contained in μ_t, and pseudo-signal variables, which are not in μ_t but are nevertheless correlated with it. We show that OCMT identifies all such covariates with probability approaching one.

In the former case where $t_{i,\mu}$ is bounded in probability sufficiently sharply as $n, T \to \infty$, we characterise x_{it} as a noise covariate if it is not contained in μ_t, and a hidden signal if it is contained in μ_t. We show that all hidden signals will be selected by the application of one or more additional stages of OCMT.

It is clear from the above exposition that our variable selection approach focusses on the net impact of x_{it} on y_t conditional on the vector of pre-selected variables $\boldsymbol{z}_t$, rather than the marginal effects defined by β_i. The conditional *net impact* coefficient of x_{it} on y_t generalizes the mean net impact coefficient considered by Pesaran and Smith (2014), and it is given by

$$\theta_{i,T}\left(\boldsymbol{z}\right) = \sum_{j=1}^{k}\beta_j\sigma_{ij,T}\left(\boldsymbol{z}\right), \tag{3}$$

where $\sigma_{ij,T}(\boldsymbol{z}) = E\left(T^{-1}\boldsymbol{x}_i'\boldsymbol{M}_z\boldsymbol{x}_j\right)$. To simplify the exposition, we suppress the T subscript and use $\theta_i\left(\boldsymbol{z}\right)$ and $\sigma_{ij}\left(\boldsymbol{z}\right)$ below.

$\theta_i(\boldsymbol{z})$ plays a crucial role in our proposed approach, as it determines whether $t_{i,\mu}$ in (2) is bounded in probability sufficiently sharply as $n, T \to \infty$. Ideally, we would like to be able to base our selection decision directly on β_i and its estimate. But when n is large such a strategy is not feasible. Instead, we propose to base variable selection on $\theta_i(\boldsymbol{z})$. It is important to stress that knowing $\theta_i(\boldsymbol{z})$ does not imply we can determine β_i. Due to the correlation between

variables, nonzero $\theta_i(\boldsymbol{z})$ does not necessarily imply nonzero β_i and we have the following four possibilities:

	$\theta_i(\boldsymbol{z}) \neq 0$	$\theta_i(\boldsymbol{z}) = 0$
$\beta_i \neq 0$	(I) Signals with nonzero net effect	(II) Hidden signals
$\beta_i = 0$	(III) Pseudo-signals	(IV) Noise variables

.

The first and the last case, where $\theta_i(\boldsymbol{z}) \neq 0$ if and only if $\beta_i \neq 0$, is the most straightforward case to be considered. But there is also a possibility of case II where $\theta_i(\boldsymbol{z}) = 0$ and $\beta_i \neq 0$ and case III where $\theta_i(\boldsymbol{z}) \neq 0$ and $\beta_i = 0$. These cases will also be considered in our analysis. Case II is likely to be rare in practice since it requires an *exact* equality between the coefficients of the true model, namely $\beta_i = -\sum_{j=1,j\neq i}^{k} \beta_j \sigma_{ii}^{-1}(\boldsymbol{z})\, \sigma_{ij}(\boldsymbol{z})$. However, the presence of pseudo-signals (case III) is quite likely, and will be an important consideration in our model selection strategy.

We shall refer to the model that contains only the signals as the *true model*, and to the model that contains the signals as well as one or more of the pseudo-signals, but none of the noise variables, as an *approximating model*. We assume that there are k^* pseudo-signal variables ordered to follow the k signal variables, so that the first $k + k^*$ variables in $\mathcal{S}_{nt}$ are signals and pseudo-signals, although this is not known to the investigator. The remaining $n - k - k^*$ variables are the noise variables. We assume that k is an unknown fixed constant, but allow k^* to rise with n such that $k^*/n \to 0$, and $k^*/T \to 0$, at a sufficiently slow rate. Specifically, we allow $k^* = \ominus(n^\epsilon)$ for some appropriately bounded $\epsilon \geq 0$. We expect ϵ to be small when the correlation between the signals and the remaining covariates is sparse.

Our secondary maintained assumptions are somewhat more general and, accordingly, lead to fewer and weaker results. A first specification assumes that there exists an ordering (possibly unknown) such that

$$\theta_i(\boldsymbol{z}) = C_i \varrho^i, \text{ for } i = 1, 2, ..., n, \text{ and } |\varrho| < 1, \tag{4}$$

for a given set of constants, C_i. A second specification modifies the decay rate and assumes that

$$\theta_i(\boldsymbol{z}) = C_i i^{-\gamma}, \text{for } i = 1, 2, ..., n, \text{ and for some } \gamma > 0. \tag{5}$$

In both specifications $\max_{1\leq i\leq n} |C_i| < C < \infty$. These specifications allow for various rates of decay in the way covariates are correlated with the signals. These cases are of technical interest and cover the autoregressive type designs considered in the literature in order to model the correlations across the covariates. See, for example, Zhang (2010) and Belloni et al. (2014b).

3 The Multiple Testing Approach

OCMT is inspired by the multiple testing literature, although the focus of OCMT is on controlling the probability of selecting an approximating model and the false discovery rate, rather

than controlling the size of the union of the multiple tests that are being carried out. To simplify the exposition below, we assume that the vector of pre-selected variables, $\boldsymbol{z}_t$, contains only an intercept, in which case, the DGP (1) simplifies to

$$y_t = a + \sum_{i=1}^{k} \beta_i x_{it} + u_t, \text{ for } t = 1, 2, ..., T. \tag{6}$$

In matrix notation, we have

$$\boldsymbol{y} = a\boldsymbol{\tau}_T + \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{u}, \tag{7}$$

where $\boldsymbol{\tau}_T$ is a $T \times 1$ vector of ones, $\boldsymbol{X}_k = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k)$ is the $T \times k$ matrix of observations on signal variables, $\boldsymbol{\beta}_k = (\beta_1, \beta_2, ..., \beta_k)'$ is the $k \times 1$ vector of associated slope coefficients and $\boldsymbol{u} = (u_1, u_2, ..., u_T)'$ is $T \times 1$ vector of errors. In addition, the conditional net impact coefficient $\theta_i(\boldsymbol{z})$ simplifies, for $\boldsymbol{z}_t = 1$, to

$$\theta_i = \sum_{j=1}^{k} \beta_j \sigma_{ij}, \tag{8}$$

where (we again suppress the subscript T), $\sigma_{ij} = E\left(T^{-1}\boldsymbol{x}_i' \boldsymbol{M}_\tau \boldsymbol{x}_j\right)$, and $\boldsymbol{M}_\tau = \boldsymbol{I}_T - \boldsymbol{\tau}_T \boldsymbol{\tau}_T'/T$. We consider the following assumptions:

Assumption 1 *Let $\boldsymbol{X}_{k,k^*} = (\boldsymbol{X}_k, \boldsymbol{X}_{k^*}^*)$, where $\boldsymbol{X}_k = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k)$, and $\boldsymbol{X}_{k^*}^* = (\boldsymbol{x}_{k+1}, \boldsymbol{x}_{k+2}, ..., \boldsymbol{x}_{k+k^*})$ are $T \times k$ and $T \times k^*$ observation matrices on signals and pseudo-signals, and suppose that there exists T_0 such that for all $T > T_0$, $\left(T^{-1}\boldsymbol{X}_{k,k^*}'\boldsymbol{X}_{k,k^*}\right)^{-1}$ is nonsingular with its smallest eigenvalue uniformly bounded away from 0, and $\boldsymbol{\Sigma}_{k,k^*} = E\left(T^{-1}\boldsymbol{X}_{k,k^*}'\boldsymbol{X}_{k,k^*}\right)$ is nonsingular for all T.*

Assumption 2 *The error term, u_t, in DGP (6) is a martingale difference process with respect to $\mathcal{F}_{t-1}^u = \sigma\left(u_{t-1}, u_{t-2}, ...,\right)$, with a zero mean and a constant variance, $0 < \sigma^2 < C < \infty$.*

Assumption 3 *Let $\mathcal{F}_{it}^x = \sigma\left(x_{it}, x_{i,t-1}, ...\right)$, where x_{it}, for $i = 1, 2, ..., n$, is the i-th covariate in the active set $\mathcal{S}_{nt}$. Define $\mathcal{F}_t^{xn} = \cup_{j=k+k^*+1}^{n} \mathcal{F}_{jt}^x$, $\mathcal{F}_t^{xo} = \cup_{i=1}^{k+k^*} \mathcal{F}_{jt}^x$, and $\mathcal{F}_t^x = \mathcal{F}_t^{xn} \cup \mathcal{F}_t^{xo}$. Then, x_{it} is independent of $x_{jt'}$ for $i = 1, 2, ..., k + k^*$, $j = k + k^* + 1, k + k^* + 2, ..., n$, and for all t and t', and $E\left[x_{it}x_{jt} - E\left(x_{it}x_{jt}\right)\middle|\mathcal{F}_{t-1}^x\right] = 0$, for $i, j = 1, 2, ..., n$, and all t. Finally, $E\left(x_{it}u_t\middle|\mathcal{F}_{t-1}\right) = 0$, for $i = 1, 2, ..., n$, and all t, where $\mathcal{F}_t = \mathcal{F}_t^x \cup \mathcal{F}_t^u$.*

Assumption 4 *There exist sufficiently large positive constants C_0, C_1, C_2 and C_3 and $s_x, s_u > 0$ such that the covariates in the active set $\mathcal{S}_{nt}$ satisfy*

$$\sup_{i,t} \Pr\left(|x_{it}| > \alpha\right) \leq C_0 \exp\left(-C_1 \alpha^{s_x}\right), \text{ for all } \alpha > 0, \tag{9}$$

and the errors, u_t, in DGP (6) satisfy

$$\sup_{t} \Pr\left(|u_t| > \alpha\right) \leq C_2 \exp\left(-C_3 \alpha^{s_u}\right), \text{ for all } \alpha > 0. \tag{10}$$

Assumption 5 *Consider x_t and the $l_T \times 1$ vector of covariates $\boldsymbol{q}_{\cdot t} = (q_{1,t}, q_{2,t}, ..., q_{l_T,t})'$. $\boldsymbol{q}_{\cdot t}$ can contain a constant term, and x_t is a generic element of $\mathcal{S}_{nt}$ that does not belong to $\boldsymbol{q}_{\cdot t}$. It is assumed that $E(\boldsymbol{q}_{\cdot t} x_t)$ and $\boldsymbol{\Sigma}_{qq} = E(\boldsymbol{q}_{\cdot t}\boldsymbol{q}'_{\cdot t})$ exist and $\boldsymbol{\Sigma}_{qq}$ is invertible. Define $\boldsymbol{\gamma}_{qx,T} = \boldsymbol{\Sigma}_{qq}^{-1}[T^{-1}\sum_{t=1}^{T} E(\boldsymbol{q}_{\cdot t} x_t)]$ and*

$$u_{x,t,T} =: u_{x,t} = x_t - \boldsymbol{\gamma}'_{qx,T}\boldsymbol{q}_{\cdot t}. \tag{11}$$

All elements of the vector of projection coefficients, $\boldsymbol{\gamma}_{qx,T}$, are uniformly bounded and only a finite number of the elements of $\boldsymbol{\gamma}_{qx,T}$ are different from zero.

Assumption 6 *The number of signals, k, in (6) is finite, and their slope coefficients could change with T, such that for $i = 1, 2, ..., k$, $\beta_{i,T} = \ominus\left(T^{-\vartheta}\right)$, for some $0 \leq \vartheta < 1/2$.*

Before formally outlining OCMT procedure and presenting our theoretical results, we provide some remarks on the pros and cons of our assumptions as compared to the ones typically assumed in the penalised regression and boosting literature.

Assumption 1 ensures that regression coefficients in the model containing all signals and pseudo-signals and none of the noise variables are identified. Assumption 2 is slightly more general than the usual assumption in the regression analysis. Assumption 3 allows x_{it} to be a martingale difference sequence which is somewhat weaker than the IID assumption typically made in the literature on penalised regression. Relaxation of this assumption to allow for serially correlated covariates is discussed in Section 4.2.

The exponential bounds in Assumption 4 are sufficient for the existence of all moments of the covariates, x_{it}, and the error term, u_t. It is very common in the literature to assume some form of exponentially declining bound for probability tails of u_t and x_{it}. See, for example, Zheng et al. (2014).

Assumption 5 is a technical condition that is required for some results derived in the Appendix and in the online theory supplement, which consider a more general multiple regression context where subsets of regressors in $\boldsymbol{x}_{nt}$ are included in the regression equation. In the simple case where $\boldsymbol{q}_{\cdot t} = 1$, then Assumption 5 is trivially satisfied and follows from the rest of the assumptions, and we have $\boldsymbol{\gamma}_{qx,T} = \mu_{x,T} = \frac{1}{T}\sum_{t=1}^{T} E(x_t)$, and $u_{x,t,T} = x_t - \mu_{x,T}$.

Assumption 6 allows for the possibility of weak signal variables whose coefficients, $\beta_{i,T}$, for $i = 1, 2, ..., k$, decline with the sample size, T, at a sufficiently slow rate. To simplify notation, subscript T is dropped subsequently, and it is understood that the slope and net effect coefficients can change with the sample size according to this assumption. Using θ_i, we can refine our concept of pseudo-signals as variables with $\theta_i = \ominus\left(T^{-\vartheta}\right)$ for $i = k+1, k+2, ..., k+k^*$, for some $0 \leq \vartheta < 1/2$. Remark 1 discusses further how this condition enters the theoretical results.

Regarding our assumptions on the correlation between variables in the active set we note the following. The signal and noise variables are allowed to be correlated amongst themselves, so no restrictions are imposed on σ_{ij} for $i, j = 1, 2, ..., k$, and on σ_{ij} for $i, j = k+k^*+1, k+k^*+2, ..., n$.

Also, signals and pseudo-signals are allowed to be correlated; namely, σ_{ij} could be non-zero for $i, j = 1, 2, ..., k + k^*$. Therefore, signals and pseudo-signals as well as noise variables can contain common factors, but, under our definition of noise variables, the factors cannot be shared between the signals/pseudo-signals and noise variables, since the latter are uncorrelated with the former. If there are common factors affecting signal variables as well as a large number of the remaining variables in the active set, one can and should condition on such factors, as we do in our empirical illustration.[4] Without such conditioning, the size of the approximating model would be too large to be of practical use, when common factors affect both signal and a large number of the remaining variables in the active set.

In contrast, a number of crucial issues arise in the context of Lasso, or more generally when L_q penalty functions with $0 \leq q \leq 1$ are used. Firstly, it is customary to assume a framework of fixed-design regressor matrices, where in many cases a generalisation to stochastic regressors is not straightforward, requiring conditions such as the spark condition of Donoho and Elad (2003) and Zheng et al. (2014). Secondly, a frequent condition for Lasso to be a valid variable selection method is the irrepresentable condition which bounds the maximum of all regression coefficients, in regression of any noise or pseudo-signal variable on the signals, to be less than one in the case of normalised regressor variables. See, for example, Section 7.5 of Buhlmann and van de Geer (2011).

Further, most results for penalised regression essentially take as given the knowledge of the tuning parameter associated with the penalty function. In practice, cross-validation is used to determine this parameter but theoretical results on the properties of such cross-validation schemes are rare. Available theoretical results on boosting, as presented in Buhlmann (2006), are also limited to the case of bounded and IID regressors, while few restrictions are placed on their correlation structure.

We proceed next with formally describing the OCMT procedure. It is a multi-stage procedure. In the first stage, we consider the n bivariate regressions of y_t on a constant ($\boldsymbol{z}_t$ in the general case) and x_{it}, for $i = 1, 2, ..., n$,

$$y_t = c_i + \phi_i x_{it} + u_{it},\ t = 1, 2, ..., T, \tag{12}$$

where $\phi_i = \theta_i / \sigma_{ii}$, θ_i is defined in (8) and σ_{ii} is defined below (8). Denoting the t-ratio of ϕ_i in this regression by $t_{\hat{\phi}_{i,(1)}}$, we have

$$t_{\hat{\phi}_{i,(1)}} = \frac{\hat{\phi}_i}{s.e.\left(\hat{\phi}_i\right)} = \frac{\boldsymbol{x}_i' \boldsymbol{M}_{\boldsymbol{\tau}} \boldsymbol{y}}{\hat{\sigma}_i \sqrt{\boldsymbol{x}_i' \boldsymbol{M}_{\boldsymbol{\tau}} \boldsymbol{x}_i}}, \tag{13}$$

[4] Note that our theory allows for conditioning on observed common factors by incorporating them in $\boldsymbol{z}_t$. But when factors are unobserved they need to be replaced by their estimates using, for example, principal components. A formal argument that the associated estimation error is asymptotically negligible involves additional technical complications, and requires deriving exponential inequalities for the quantities analysed in Theorem 1 of Bai and Ng (2002) and Lemma A1 of Bai and Ng (2006), and then assuming that $\sqrt{T}/n \to 0$ as $n, T \to \infty$. While such a derivation is clearly feasible under appropriate regularity conditions, a formal analysis is beyond the scope of the present paper.

where $\hat{\phi}_i = (\boldsymbol{x}_i' \boldsymbol{M}_{\boldsymbol{\tau}} \boldsymbol{x}_i)^{-1} \boldsymbol{x}_i' \boldsymbol{M}_{\boldsymbol{\tau}} \boldsymbol{y}$ denotes the LS estimator of ϕ_i, $\hat{\sigma}_i^2 = \boldsymbol{e}_i' \boldsymbol{e}_i / T$, and $\boldsymbol{e}_i$ denotes the $T \times 1$ vector of residual of the regression of $\boldsymbol{y}$ on $\boldsymbol{\tau}_T$ and $\boldsymbol{x}_i$. The first stage OCMT selection indicator is given by

$$\widehat{\mathcal{J}}_{i,(1)} = I[|t_{\hat{\phi}_{i,(1)}}| > c_p(n, \delta)], \text{ for } i = 1, 2, ..., n, \tag{14}$$

where $c_p(n, \delta)$ is a *critical value function* defined by

$$c_p(n, \delta) = \Phi^{-1}\left(1 - \frac{p}{2f(n, \delta)}\right), \tag{15}$$

$\Phi^{-1}(.)$ is the inverse of standard normal distribution function, $f(n, \delta) = cn^{\delta}$ for some positive constants δ and c, and p $(0 < p < 1)$ is the nominal size of the individual tests to be set by the investigator. We will refer to δ as the *critical value exponent*. One value of δ is used in the first stage, while another one (denoted by δ^*) is used in subsequent stages of OCMT. As we shall see, it will be required that $\delta^* > \delta$. Variables with $\widehat{\mathcal{J}}_{i,(1)} = 1$ are selected as signals and pseudo-signals in the first stage. Denote the number of covariates selected in the first stage by $\hat{k}_{(1)}^o$, the index set of the selected variables by $\mathcal{S}_{(1)}^o$, and the $T \times \hat{k}_{(1)}^o$ observation matrix of the $\hat{k}_{(1)}^o$ selected variables by $\boldsymbol{X}_{(1)}^o$. Further, let $\boldsymbol{X}_{(1)} = (\boldsymbol{\tau}_T, \boldsymbol{X}_{(1)}^o) = (\boldsymbol{x}_{(1),1}, ..., \boldsymbol{x}_{(1),T})'$, $\hat{k}_{(1)} = \hat{k}_{(1)}^o$, $\mathcal{S}_{(1)} = \mathcal{S}_{(1)}^o$, and $\mathfrak{A}_{(2)} = \{1, 2, ..., n\} \setminus \mathcal{S}_{(1)}$. For future reference, we also set $\boldsymbol{X}_{(0)} = \boldsymbol{\tau}_T$ and $\mathfrak{A}_{(1)} = \{1, 2, ..., n\}$. In stages $j = 2, 3, ...$, we consider the $n - \hat{k}_{(j-1)}$ regressions of y_t on the variables in $\boldsymbol{X}_{(j-1)}$ and, one at the time, x_{it} for i belonging in the active set, $\mathfrak{A}_{(j)}$. We then compute the following t-ratios

$$t_{\hat{\phi}_{i,(j)}} = \frac{\hat{\phi}_{i,(j)}}{s.e.\left(\hat{\phi}_{i,(j)}\right)} = \frac{\boldsymbol{x}_i' \boldsymbol{M}_{(j-1)} \boldsymbol{y}}{\hat{\sigma}_{i,(j)} \sqrt{\boldsymbol{x}_i' \boldsymbol{M}_{(j-1)} \boldsymbol{x}_i}}, \text{ for } i \in \mathfrak{A}_{(j)}, \; j = 2, 3, ..., \tag{16}$$

where $\hat{\phi}_{i,(j)} = \left(\boldsymbol{x}_i' \boldsymbol{M}_{(j-1)} \boldsymbol{x}_i\right)^{-1} \boldsymbol{x}_i' \boldsymbol{M}_{(j-1)} \boldsymbol{y}$ is the LS estimator of the conditional net effect of x_{it} on y_t in stage j, $\hat{\sigma}_{i,(j)}^2 = T^{-1} \boldsymbol{e}_{i,(j)}' \boldsymbol{e}_{i,(j)}$, $\boldsymbol{M}_{(j-1)} = \boldsymbol{I}_T - \boldsymbol{X}_{(j-1)} (\boldsymbol{X}_{(j-1)}' \boldsymbol{X}_{(j-1)})^{-1} \boldsymbol{X}_{(j-1)}'$, and $\boldsymbol{e}_{i,(j)}$ denotes the residual vector of the regression of $\boldsymbol{y}$ on $\boldsymbol{X}_{i,(j-1)} = \left(\boldsymbol{x}_i, \boldsymbol{X}_{(j-1)}\right)$. Regressors for which $\widehat{\mathcal{J}}_{i,(j)} = 1$, are then added to the set of already selected covariates from the previous stages, where $\widehat{\mathcal{J}}_{i,(j)} = I[|t_{\hat{\phi}_{i,(j)}}| > c_p(n, \delta^*)]$. Denote the number of variables selected in stage j by $\hat{k}_{(j)}^o$, their index set by $\mathcal{S}_{(j)}^o$, and the $T \times \hat{k}_{(j)}^o$ matrix of the $\hat{k}_{(j)}^o$ selected variables in stage j by $\mathbf{X}_{(j)}^o$. Also let $\mathbf{X}_{(j)} = (\mathbf{X}_{(j-1)}, \mathbf{X}_{(j)}^o) = (\boldsymbol{x}_{(j),1}, \boldsymbol{x}_{(j),2}, ..., \boldsymbol{x}_{(j),T})'$, $\hat{k}_{(j)} = \hat{k}_{(j-1)} + \hat{k}_{(j)}^o$, $\mathcal{S}_{(j)} = \mathcal{S}_{(j-1)} \cup \mathcal{S}_{(j)}^o$, define the $(j+1)$ stage active set by $\mathfrak{A}_{(j+1)} = \{1, 2, ..., n\} \setminus \mathcal{S}_{(j)}$, and then proceed to the next stage by increasing j by one. Note that $\hat{k}_{(j)}$ is the *total* number of variables selected up to and including stage j, $\hat{\phi}_{i,(j)} \to_p \theta_{i,(j)} / \sigma_{ii,(j)}$, where $\theta_{i,(j)}$ and $\sigma_{ii,(j)}$ are used in the remainder of this paper to denote $\theta_i\left(\boldsymbol{x}_{(j-1)}\right)$ and $\sigma_{ii}\left(\boldsymbol{x}_{(j-1)}\right)$ introduced in (3). Also to simplify the notation, $\theta_{i,(1)}$ is shown as θ_i. The procedure stops when no regressors are selected at a given stage, say $\hat{\jmath}$, in which case the final number of selected variables will be given, as before,

by $\hat{k} = \hat{k}_{(\hat{\jmath}-1)}$. The multi-stage OCMT selection indicator is thus given by $\widehat{\mathcal{J}}_i = \sum_{j=1}^{\hat{P}} \widehat{\mathcal{J}}_{i,(j)}$, where $\hat{P}$ denotes the number of stages at completion of OCMT, formally defined as

$$\hat{P} = \min_j \{j : \textstyle\sum_{i=1}^{n} \widehat{\mathcal{J}}_{i,(j)} = 0\} - 1. \tag{17}$$

It is important to note that the number of stages needed for OCMT is bounded in n. To show this we note that not all signals can be hidden, and once we condition on the set of signals that are not hidden, then there must exist i such that $\theta_i(\boldsymbol{z}) \neq 0$, while $\theta_i = 0$ and $\beta_i \neq 0$, where here $\boldsymbol{z}$ denotes the signal variables that are not hidden.[5] Using this result one can successively uncover all hidden signals. We denote by P the number of stages that need to be considered to uncover all hidden signals. Its true population value is denoted by P_0. This is defined as the index of the last stage where OCMT finds further signals (or pseudo-signals), assuming that $\Pr[|t_{\hat{\phi}_{i,(j)}}| > c_p(n,\delta) \,|\theta_{i,(j)} \neq 0] = 1$ and $\Pr[|t_{\hat{\phi}_{i,(j)}}| > c_p(n,\delta) \,|\theta_{i,(j)} = 0] = 0$, for all variables indexed by i, and OCMT stages indexed by j. Of course, these probabilities do not take the values 1 and 0 respectively, in small samples, but we will handle this complication later on. The following proposition provides an upper bound to P_0.

Proposition 1 *Suppose that y_t, $t = 1, 2, ..., T$, are generated according to (6), with $\beta_i \neq 0$ for $i = 1, 2, ..., k$, and that Assumption 1 holds. Then, there exists j, $1 \leq j \leq k$, for which $\theta_{i,(j)} \neq 0$, and the population value of the number of stages required to select all the signals, denoted as P_0, satisfies $1 \leq P_0 \leq k$.*

A proof is provided in Subsection A.2.1 of the Appendix.

In practice, $\hat{P}$ is likely to be small since hidden signals arise only in rare cases where $\theta_i = 0$ whilst the associated β_i is non-zero. Also, as we show all signals with nonzero θ will be picked up with probability tending to one in the first stage. Stopping after the first stage tends to improve the small sample performance of the OCMT approach, investigated in Section 5, only marginally when no hidden signals are present. Thus, allowing $P > 1$, using the stopping rule defined above, does not significantly deteriorate the small sample performance of OCMT when hidden signals are not present, while it picks-up all hidden signals with probability tending to one. Finally, using (7), note that the conditional net effect coefficient of variable i at stage j of OCMT, $\theta_{i,(j)}$, can be written as

$$\theta_{i,(j)} = E\left(T^{-1}\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\boldsymbol{y}\right) = E\left(T^{-1}\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\boldsymbol{X}_k\boldsymbol{\beta}_k\right) = \textstyle\sum_{\ell=1}^{k}\beta_\ell\sigma_{i\ell}\left(\boldsymbol{x}_{(j-1)}\right), \tag{18}$$

and to allow for the possibility of weak signals as defined by Assumption 6, pseudo-signal variables can be more generally defined as covariates $i = k+1, k+2, ..., k+k^*$ with $\theta_{i,(j)} = \ominus\left(T^{-\vartheta}\right)$, for some $0 \leq \vartheta < 1/2$ and some $1 \leq j \leq P_0$.

[5] For a proof see Lemma A1 in the online supplement. Note also that $\boldsymbol{z}_t$ may contain lagged values of y_t, principal components or other estimates of common effects as well as covariates that the investigator believes must be included.

Once the OCMT procedure is completed, the OCMT estimator of β_i, denoted by $\tilde{\beta}_i$, is set as

$$\tilde{\beta}_i = \begin{cases} \hat{\beta}_i^{(\hat{k})}, & \text{if } \widehat{\mathcal{J}}_i = 1 \\ 0, & \text{otherwise} \end{cases}, \text{ for } i = 1, 2, ..., n, \tag{19}$$

where $\hat{\beta}_i^{(\hat{k})}$ is the LS estimator of the coefficient of the i^{th} variable in a regression of y_t on all the selected covariates, namely *all* the covariates for which $\widehat{\mathcal{J}}_i = 1$, plus a constant term ($\boldsymbol{z}_t$ in the general case).

The choice of the critical value function, $c_p(n, \delta)$, given by (15), is important since it allows the investigator to relate the size and power of the selection procedure to the inferential problem in classical statistics, with the modification that p (type I error) is now scaled by a function of the number of covariates under consideration. As we shall see, the OCMT procedure applies irrespective of whether n is small or large relative to T, so long as $T = \ominus(n^{\kappa_1})$, for any finite $\kappa_1 > 0$. This follows from result (i) of Lemma A2 in the online supplement, which establishes that $c_p^2(n, \delta) = O[\delta \ln(n)]$. It is also helpful to bear in mind that, using result (ii) of Lemma A2 in the online supplement, $\exp\left[-\varkappa c_p^2(n, \delta)/2\right] = \ominus\left(n^{-\delta\varkappa}\right)$, and $c_p(n, \delta) = o\left(T^{C_0}\right)$, for all $C_0 > 0$, assuming there exists $\kappa_1 > 0$, such that $T = \ominus(n^{\kappa_1})$.

Note that setting $\delta = 1$ in the first stage, is equivalent to using a Bonferroni correction for the multiple testing problem. Of course, other c_p values can be used, such as those proposed by Holm (1979), Benjamini and Hochberg (1995), or Gavrilov et al. (2009) which are designed to control the family-wise error rate associated with a set of tests. However, since most impose some restriction on the dependence structure between the multiple tests (with the exception of the original Bonferroni procedure and the one proposed by Holm (1979)), we choose to use (15) which, furthermore, has a bespoke design, in terms of the conditions placed on δ, and is appropriate for the multi-stage OCMT method, where the number of tests carried out is not predetermined in advance.

We now consider the relationship of OCMT to sequential model selection procedures advanced in the literature. A notable example is L_2-Boosting by Buhlmann (2006) which starts with the same set of bivariate regressions, (12), but in the first step selects *only* the covariate with the maximum fit, as measured by the sum of squared residuals (SSR). Additional covariates are added sequentially by regressing a quasi-residual from the first step on the remaining covariates. The process is continued till convergence decided based on some information criterion.[6] Other sequential model selection approaches, such as those by Fithian et al. (2014), Tibshirani et al. (2014) and Fithian et al. (2015) build regression models by selecting variables from active sets, based on a sequence of tests. Variables are selected, and added to the model, one by one and selection stops once a test does not reject the latest null hypothesis in the sequence. It is important to note that these methods select one covariate (or at most a block

[6] The quasi-residuals are computed as $y_t - v\,\hat{y}_t$, where $\hat{y}_t$ is the fitted value in terms of the selected covariate, and v is a constant tuning parameter referred to as the step size. Buhlmann (2006) recommends choosing $v < 1$.

of covariates) in each of the steps. In contrast, OCMT operates as a 'hub and spoke' approach. It selects, in a single step, all variables whose t-ratios, in (12), exceed a threshold (given by $c_p(n,\delta)$), in absolute value. As a result, it is clear that in its main implementation OCMT is not a sequential approach. Only in the presence of hidden signals, does OCMT require subsequent stages. Even then, under our setting, where k is finite, the number of stages cannot exceed k with a high probability, and as a result in the vast majority of cases the number of additional stages required will be rather small.

We investigate the asymptotic properties of the OCMT procedure and the associated OCMT estimators, $\tilde{\beta}_i$, for $i = 1, 2, ..., n$, in terms of the probability of selecting the approximating model, and in terms of support recovery type statistics used in the Lasso literature, namely the true and false positive rates (TRP and FPR, respectively) defined by

$$TPR_{n,T} = \frac{\sum_{i=1}^{n} I(\widehat{\mathcal{J}}_i = 1 \text{ and } \beta_i \neq 0)}{\sum_{i=1}^{n} I(\beta_i \neq 0)}, \text{ and } FPR_{n,T} = \frac{\sum_{i=1}^{n} I(\widehat{\mathcal{J}}_i = 1, \text{ and } \beta_i = 0)}{\sum_{i=1}^{n} I(\beta_i = 0)}. \quad (20)$$

We also examine the following false discovery rate

$$FDR_{n,T} = \frac{\sum_{i=1}^{n} I(\widehat{\mathcal{J}}_i = 1, \text{ and } \beta_i = \theta_i = 0)}{\sum_{i=1}^{n} \widehat{\mathcal{J}}_i + 1}, \quad (21)$$

which applies to selection of signals and pseudo-signals. Further, we consider the error and the coefficient norms of the selected model, defined by

$$F_{\tilde{u}} = T^{-1}||\tilde{\boldsymbol{u}}||^2 = T^{-1}\sum_{t=1}^{T}\tilde{u}_t^2, \text{ and } F_{\tilde{\boldsymbol{\beta}}} = ||\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n|| = [\sum_{i=1}^{n}(\tilde{\beta}_i - \beta_i)^2]^{1/2}, \quad (22)$$

respectively, where $\tilde{\boldsymbol{u}} = (\tilde{u}_1, \tilde{u}_2, ..., \tilde{u}_T)'$, $\tilde{u}_t = y_t - \hat{a} - \tilde{\boldsymbol{\beta}}_n' \boldsymbol{x}_{nt}$, $\boldsymbol{\beta}_n = (\beta_1, \beta_2, ..., \beta_n)'$, $\tilde{\boldsymbol{\beta}}_n = (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_n)'$, $\tilde{\beta}_i$, for $i = 1, 2, ..., n$ are defined by (19), and $\hat{a}$ is the estimator of the constant term in the final regression.

We now present the main theoretical results using lemmas established in the online supplement. The key is Lemma A10 in the online supplement, which provides sharp bounds on the probability of $|t_{\hat{\phi}_{i,(j)}}| > c_p(n,\delta)$ conditional on whether the net effect coefficient $\theta_{i,(j)}$ is zero or not. Here we provide a simpler version of this lemma which focuses on the first-stage regressions and should provide a better understanding of the main mathematical results that lie behind the proofs in the more complicated multi-stage version of the OCMT.

Proposition 2 *Suppose y_t is given by (6) and Assumptions 2-4 hold. Let x_t be a generic element of the active set $\mathcal{S}_{nt}$, and suppose Assumption 5 holds for x_t and $\boldsymbol{q}_{.t} = 1$. Consider the t-ratio of x_t in the regression of y_t on an intercept and x_t:*

$$t_x = \frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_\tau \boldsymbol{y}}{\sqrt{(T^{-1}\boldsymbol{e}'\boldsymbol{e})(T^{-1}\boldsymbol{x}'\mathbf{M}_\tau \boldsymbol{x})}},$$

where $\boldsymbol{e}$ is the $T \times 1$ vector of regression residuals. Let $\theta = E(T^{-1}\boldsymbol{x}'\mathbf{M}_\tau \boldsymbol{y})$ be the net impact effect of x_t, and suppose there exists $\kappa_1 > 0$ such that $T = \ominus(n^{\kappa_1})$. Then, for some finite

positive constants C_0 and C_1, we have

$$\Pr\left[|t_x| > c_p\left(n,\delta\right)|\theta = 0\right] \leq \exp\left[-\chi c_p^2\left(n,\delta\right)/2\right] + \exp\left(-C_0 T^{C_1}\right), \tag{23}$$

where $c_p\left(n,\delta\right)$ is the critical value function given by (15), and $\chi = \left[\left(1-\pi\right)/\left(1+d_T\right)\right]^2$, for any π in the range $0<\pi<1$, any $d_T > 0$ and bounded in T. Suppose further that in the case where $\theta \neq 0$, we have $\theta = \ominus\left(T^{-\vartheta}\right)$, for some $0 \leq \vartheta < 1/2$, where $c_p\left(n,\delta\right) = O\left(T^{1/2-\vartheta-C_4}\right)$, for some positive constant C_4. Then,

$$\Pr\left[|t_x| > c_p\left(n,\delta\right)|\theta \neq 0\right] > 1 - \exp\left(-C_2 T^{C_3}\right). \tag{24}$$

Result (23) establishes a sharp probability bound for the absolute value of the t-ratio of x with zero net impact effect. The first term on the right side of (23) asymptotically dominates, and using result (ii) of Lemma A2 in the online supplement we have $\exp\left[-\chi c_p^2\left(n,\delta\right)/2\right] = \ominus\left(n^{-\delta\chi}\right)$. Result (24), on the other hand, establishes a lower bound on the probability of the event $|t_{\hat{\phi}_{i,(1)}}| > c_p\left(n,\delta\right)$ conditional on θ being sufficiently away from zero.

Since we wish to allow for the possibility of hidden signals for which $\theta = 0$ even if the associated $\beta \neq 0$, the results in Lemma A10 in the online supplement are obtained for t-ratios in multiple regression contexts where subsets of regressors in the active set are also included in the regression equation for y_t. Nevertheless, it is instructive to initially consider the OCMT in the absence of such hidden signals. Theorems 1 and 2 below provide the results for the general case where hidden signals are allowed.

We first examine $TPR_{n,T}$ defined by (20), under the assumption that $\theta_i \neq 0$ if $\beta_i \neq 0$. Note that by definition $TPR_{n,T} = k^{-1}\sum_{i=1}^{k} I(\widehat{\mathcal{J}}_{i,(1)} = 1 \text{ and } \beta_i \neq 0)$. Since the elements of this summation are 0 or 1, then taking expectations we have (note that in the present simple case $\theta_i \neq 0$ implies $\beta_i \neq 0$)

$$TPR_{n,T} = k^{-1}\textstyle\sum_{i=1}^{k} E[I(\widehat{\mathcal{J}}_{i,(1)} = 1 \text{ and } \beta_i \neq 0)] = k^{-1}\sum_{i=1}^{k} \Pr[|t_{\hat{\phi}_{i,(1)}}| > c_p\left(n,\delta\right)|\theta_i \neq 0].$$

Now using result (24) of Proposition 2, and recalling that $T = \ominus\left(n^{\kappa_1}\right)$, we have

$$TPR_{n,T} \geq 1 - \exp\left(-C_2 T^{C_3}\right) = 1 + O\left[\exp\left(-C_2 n^{C_3\kappa_1}\right)\right], \tag{25}$$

for some $C_2, C_3 > 0$. Hence, $TPR_{n,T} \to_p 1$ for any $\kappa_1 > 0$.

Consider now $FPR_{n,T}$ defined by (20). Again, note that the elements of $FPR_{n,T}$ are either 0 or 1 and hence $|FPR_{n,T}| = FPR_{n,T}$. Taking expectations of the right part of (20), and assuming $\theta_i = \ominus\left(T^{-\vartheta}\right)$, for $i = k+1, k+2, ..., k+k$, and some $0 \leq \vartheta < 1/2$, we have $\left(n-k\right)^{-1}\sum_{i=k+1}^{n} \Pr[|t_{\hat{\phi}_{i,(1)}}| > c_p\left(n,\delta\right)|\beta_i = 0] = \left(n-k\right)^{-1}\sum_{i=k+1}^{k+k^*} \Pr[|t_{\hat{\phi}_{i,(1)}}| > c_p\left(n,\delta\right)|\theta_i \neq 0] + \left(n-k\right)^{-1}\sum_{i=k+k^*+1}^{n} \Pr[|t_{\hat{\phi}_{i,(1)}}| > c_p\left(n,\delta\right)|\theta_i] = 0$. Using (24) of Proposition 2 and assuming there exists $\kappa_1 > 0$ such that $T = \ominus\left(n^{\kappa_1}\right)$, we have $k^* - \sum_{i=k+1}^{k+k^*} \Pr[|t_{\hat{\phi}_{i,(1)}}| > c_p\left(n,\delta\right)|\theta_i \neq 0] = O\left[\exp\left(-C_2 T^{C_3}\right)\right]$, for some finite positive constants C_2 and C_3. Moreover, (23) of Proposition

2, which holds uniformly over i, given the uniformity of (9) and (10) of Assumption 4, implies that for any $0 < \varkappa < 1$ there exist finite positive constants C_0 and C_1 such that

$$\sum_{i=k+k^*+1}^{n} \Pr[|t_{\hat{\phi}_{i,(1)}}| > c_p(n,\delta) \,|\theta_i = 0] \leq \sum_{i=k+k^*+1}^{n} \left\{ \exp\left[-\varkappa c_p^2(n,\delta)/2\right] + \exp\left(-C_0 T^{C_1}\right) \right\}. \tag{26}$$

Using these results we obtain

$$\begin{aligned}(n-k)^{-1} \sum_{i=k+1}^{n} \Pr[|t_{\hat{\phi}_{i,(1)}}| > c_p(n,\delta) \,|\beta_i = 0] = k^*/(n-k) + O\left\{\exp\left[-\varkappa c_p^2(n,\delta)/2\right]\right\} \\ + O\left[\exp(-C_0 T^{C_1})\right] + O\left[\exp\left(-C_2 T^{C_3}\right)/(n-k)\right]. \end{aligned} \tag{27}$$

Next, we consider the probability of choosing the approximating model. A selected regression model is referred to as an approximating model if it contains the signal variables x_{it}, $i = 1, 2, ..., k$, and none of the noise variables, x_{it}, $i = k + k^* + 1, k + k^* + 2, ..., n$. The models in the set may contain one or more of the pseudo-signals, x_{it}, $i = k + 1, k + 2, ..., k + k^*$. We refer to all such regressions as the set of approximating models. So, the event of choosing the approximating model is given by

$$\mathcal{A}_0 = \{\textstyle\sum_{i=1}^{k} \widehat{\mathcal{J}}_i = k\} \cap \{\sum_{i=k+k^*+1}^{n} \widehat{\mathcal{J}}_i = 0\}. \tag{28}$$

Theorem 1 below states the conditions under which $\Pr(\mathcal{A}_0) \to 1$. The results for the general multi-stage case that allows for the possibility of hidden signals are given in the following theorem. Since it is assumed that the expansion rates of T and n are related, the results that follow are reported in terms of n for presentational ease and consistency. They could, of course, be reported equally in terms of T, if required.

Theorem 1 *Consider the DGP (6) with k signals, k^* pseudo-signals, and $n - k - k^*$ noise variables, and suppose that Assumptions 1-4 and 6 hold, Assumption 5 holds for x_{it} and $\boldsymbol{q}_{.t} = \boldsymbol{x}_{(j-1),t}$, $i \in \mathfrak{A}_{(j)}$, $j = 1, 2, ...k$, where $\mathfrak{A}_{(j)}$ is the active set at stage j of the OCMT procedure. $c_p(n,\delta)$ is given by (15) with $0 < p < 1$ and let $f(n,\delta) = cn^{\delta}$, for the first stage of OCMT, and $f(n,\delta^*) = cn^{\delta^*}$ for subsequent stages, for some $c > 0$, $\delta^* > \delta > 0$. $n, T \to \infty$, such that $T = \ominus(n^{\kappa_1})$, for some $\kappa_1 > 0$, and $k^* = \ominus(n^{\epsilon})$ for some positive $\epsilon < \min\{1, \kappa_1/3\}$. Then, for any $0 < \varkappa < 1$, and for some constant $C_0 > 0$,*

(a) the probability that the number of stages in the OCMT procedure, $\hat{P}$, defined by (17), exceeds k is given by

$$\Pr\left(\hat{P} > k\right) = O\left(n^{1-\varkappa\delta^*}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[\exp\left(-n^{C_0\kappa_1}\right)\right], \tag{29}$$

(b) the probability of selecting the approximating model, $\mathcal{A}_0$, defined by (28), is given by

$$\Pr(\mathcal{A}_0) = 1 + O\left(n^{1-\delta\varkappa}\right) + O\left(n^{2-\delta^*\varkappa}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[\exp\left(-n^{C_0\kappa_1}\right)\right], \tag{30}$$

(c) for the True Positive Rate, $TPR_{n,T}$, defined by (20), we have

$$E\left|TPR_{n,T}\right| = 1 + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[\exp\left(-n^{C_0\kappa_1}\right)\right], \tag{31}$$

and if $\delta > 1 - \kappa_1/3$, then $TPR_{n,T} \to_p 1$; for the False Positive Rate, $FPR_{n,T}$, defined by (20), we have

$$E\left|FPR_{n,T}\right| = \frac{k^*}{n-k} + O\left(n^{-\varkappa\delta}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left(n^{\epsilon-1}\right) + O\left[\exp\left(-n^{C_0\kappa_1}\right)\right], \tag{32}$$

and if $\delta > \min\{0, 1-\kappa_1/3\}$, and $\delta^* > 1$, then $FPR_{n,T} \to_p 0$. For the False Discovery Rate, $FDR_{n,T}$, defined in (21), we have $FDR_{n,T} \to_p 0$, if $\delta > \max\{1, 2-\kappa_1/3\}$.

Since our proof requires that $0 < \varkappa < 1$, it is sufficient to set $\varkappa$ to be arbitrarily close to, but less than, unity. Also, κ_1 can be arbitrarily small which allows n to rise much faster than T. The condition $0 \leq \epsilon < \min\{1, \kappa_1/3\}$ ensures that $k^*/n \to 0$ and $k^* = o(T^{1/3})$.

Remark 1 *Assumption 6 allows for weak signals. In particular, we allow slope coefficients of order $\ominus\left(T^{-\vartheta}\right)$, for some $0 \leq \vartheta < 1/2$. Then, by (B.57) and (B.58) of Lemma A10 of the online supplement, it is seen that such weak signals can be picked up at no cost, in terms of rates, with respect to the exponential inequalities that underlie all the theoretical results. In particular, the power of the OCMT procedure in selecting the signal variable x_{it} rises with the ratio $\sqrt{T}\left|\theta_{i,(j)}\right| / \sigma_{e_i,(T)}\sigma_{x_i,(T)}$, so long as $\frac{c_p(n,\delta)}{\sqrt{T}\left|\theta_{i,(j)}\right|} \to 0$, as n and $T \to \infty$, where $\theta_{i,(j)}$ is given by (18), $\sigma_{e_i,(T)}$ and $\sigma_{x_i,(T)}$ are defined by (B.49), replacing $\boldsymbol{e}$, $\boldsymbol{x}$, and $\boldsymbol{M}_q$ by $\boldsymbol{e}_i$, $\boldsymbol{x}_i$, and $\boldsymbol{M}_{(j-1)}$, respectively. When this ratio is low, a large T will be required for the OCMT approach to select the i^{th} signal variable. This condition is similar to the so-called 'beta-min' condition assumed in the penalised regression literature. (See, for example, Section 7.4 of Buhlmann and van de Geer (2011) for a discussion.)*

Remark 2 *When the focus of the analysis is the true model, and not the approximating model that encompasses it, then the false discovery rate of the true model is given by*

$$FDR^*_{n,T} = \frac{\sum_{i=1}^n I(\widehat{\mathcal{J}}_i = 1, \text{ and } \beta_i = 0)}{\sum_{i=1}^n \widehat{\mathcal{J}}_i + 1}. \tag{33}$$

It is now easily seen that $FDR^*_{n,T}$ can tend to a nonzero value when pseudo-signals are present (i.e. if $k^* > 0$). In such cases, where the selection of the true model is the main objective of the analysis, a post-OCMT selection, using, for example, the Schwarz information criterion, could be considered to separate the signals from the pseudo-signals. However, when the norm of slope coefficients or the in-sample fit of the model is of main concern, then, under appropriate conditions on the rate at which k^* expands with n, the inclusion of pseudo-signals is asymptotically innocuous, as shown in Theorem 2 below.

Consider now the error and coefficient norms of the selected model, $F_{\boldsymbol{u}}$ and $F_{\tilde{\boldsymbol{\beta}}}$, defined in (22). We need the following additional regularity condition.

Assumption 7 *Let* $\mathbf{S}$ *denote the* $T \times l_T$ *observation matrix on the* l_T *regressors selected by the OCMT procedure. Then, let* $\mathbf{\Sigma}_{ss} = E\left(\mathbf{S}'\mathbf{S}/T\right)$ *with eigenvalues denoted by* $\mu_1 \leq \mu_2 \leq ... \leq \mu_{l_T}$. *Let* $\mu_i = O\left(l_T\right)$, $i = l_T - M + 1, l_T - M + 2, ..., l_T$, *for some finite* M, *and* $\sup_{1 \leq i \leq l_T - M} \mu_i < C_0 < \infty$, *for some* $C_0 > 0$. *In addition,* $\inf_{1 \leq i < l_T} \mu_i > C_1 > 0$, *for some* $C_1 > 0$.

Theorem 2 *Consider the DGP defined by (6), and the error and coefficient norms of the selected model,* $F_{\tilde{u}}$ *and* $F_{\tilde{\boldsymbol{\beta}}}$, *defined in (22). Suppose that Assumptions 1-4 and 6-7 hold, Assumption 5 holds for* x_{it} *and* $\boldsymbol{q}_{.t} = \boldsymbol{x}_{(j-1),t}$, $i \in \mathfrak{A}_{(j)}$, $j = 1, 2, ...k$, *where* $\mathfrak{A}_{(j)}$ *is the active set at stage* j *of the OCMT procedure, and* k^* *(the number of pseudo-signals) is of order* $\ominus\left(n^{\epsilon}\right)$ *for some positive* ϵ. $c_p\left(n, \delta\right)$ *is given by (15) with* $0 < p < 1$ *and let* $f\left(n, \delta\right) = cn^{\delta}$, *for the first stage of OCMT, and* $f\left(n, \delta^*\right) = cn^{\delta^*}$ *for subsequent stages, for some* $c > 0$, $\delta^* > \delta > 0$. $n, T \to \infty$, *such that* $T = \ominus\left(n^{\kappa_1}\right)$, *for some* $\kappa_1 > 0$, *and* $k^* = \ominus(n^{\epsilon})$ *for some positive* $\epsilon < \min\left\{1, \kappa_1/3\right\}$. *Let* $\tilde{\boldsymbol{\beta}}_n$ *be the estimator of* $\boldsymbol{\beta}_n = (\beta_1, \beta_2, ..., \beta_n)'$ *in the final regression. Then, for any* $0 < \varkappa < 1$, *and some constant* $C_0 > 0$, *we have*

$$F_{\tilde{u}} = T^{-1}||\tilde{\boldsymbol{u}}||^2 = \sigma^2 + O_p(T^{-1/2}) + O(n^{3\epsilon}T^{-3/2}) = \sigma^2 + O_p(n^{-\kappa_1/2}) + O(n^{3\epsilon - 3\kappa_1/2}), \tag{34}$$

and

$$F_{\tilde{\boldsymbol{\beta}}} = ||\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n|| = O_p(n^{5\epsilon/2}T^{-1}) = O_p(n^{5\epsilon/2 - \kappa_1}). \tag{35}$$

As can be seen from the above theorem, (34) and (35) require slightly stronger conditions than those needed for the proof of the earlier results in Theorem 1. In particular, a condition that relates to the eigenvalues of the population covariance of the selected regressors, denoted by $\mathbf{\Sigma}_{ss}$, is needed. It aims to control the rate at which $\left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F$ grows. It is mild in the sense that it allows for the presence of considerable collinearity between the regressors. Under this condition and $\epsilon < \min\left\{1, \kappa_1/3\right\}$, we in fact obtain an oracle rate of $T^{-1/2}$ for the error norm.

It is important to provide intuition on why we can get a consistency result for the coefficient norm of the selected model even though the selection process includes pseudo-signals. There are two reasons for this. First, since OCMT procedure selects all signals with probability approaching one as $n, T \to \infty$, then the coefficients of the additionally selected regressors (whether pseudo-signal or noise) will tend to zero with T. Second, restricting the rate at which k^* rises with n, as set out in Theorem 2, implies that the inclusion of pseudo-signals can be accommodated since their estimated coefficients will tend to zero and the variance of these estimated coefficients will be controlled.

In the case where hidden signals are not present, we have $P_0 = 1$, and as noted earlier further stages of the OCMT will not be required. Consequently, the results of Theorem 1 can be simplified and obtained under a less restrictive set of conditions. When $P_0 = 1$, and assuming that the conditions of Theorem 1 hold, with the exception of the condition on ϵ which could lie in $[0, 1)$, we obtain the following results, established in Section A.2.5 of the Appendix. The probability of selecting the approximating model is given by

$$\Pr\left(\mathcal{A}_0\right) = 1 + O\left(n^{1-\delta\varkappa}\right) + O\left[n\exp\left(-n^{C_0}\right)\right], \tag{36}$$

and $\Pr(\mathcal{A}_0) \to_p 1$, if $\delta > 1$. For the support recovery statistics, we have

$$E|TPR_{n,T}| = 1 + O\left[\exp\left(-n^{C_0}\right)\right], \text{ and} \tag{37}$$

$$E|FPR_{n,T}| = k^*/(n-k) + O\left(n^{-\delta\varkappa}\right) + O\left(n^{\epsilon-1}\right) + O\left[\exp(-n^{C_0})\right]. \tag{38}$$

Hence, if $\delta > 0$, then $TPR_{n,T} \to_p 1$, and $FPR_{n,T} \to_p 0$, and $FDR_{n,T} \to_p 0$, if $\delta > 1$.

4 Extensions

4.1 Alternative specifications for θ_i

Theorems 1 and 2, and the results discussed above relate to the first maintained assumption about the pseudo-signal variables where at most k^* of them have non-zero $\theta_{i,(j)}$ for some j. This result can be extended to the case where potentially all variables have non-zero θ_i, as long as θ_i's are absolutely summable. Two leading cases considered in the literature are to assume that there exists a (possibly unknown) ordering given by (4) or (5). The assumption that there is only a finite number of variables for which $\beta_i \neq 0$, is retained. The rationale for hidden signals is less clear for these cases, since rather than a discrete separation between variables with zero and non-zero θ_i, we consider a continuum that unites these two classes of variables. Essentially, we have no separation in terms of signals (or pseudo-signals) and noise variables, since under this setting there are no noise variables. Below, we provide some results for the settings implied by (4) and (5), proven in the online supplement.

Theorem 3 *Consider the DGP defined by (6), suppose that Assumptions 1-4 and 6 hold, Assumption 5 holds for x_{it} and $\boldsymbol{q}_{\cdot t} = 1$, $i = 1, 2, ..., n$, and condition (4) holds. Moreover, let $c_p(n, \delta)$ be given by (15) with $0 < p < 1$ and $f(n, \delta) = cn^{\delta}$, for some $c, \delta > 0$, and suppose there exists $\kappa_1 > 0$ such that $T = \ominus(n^{\kappa_1})$. Consider the variables selected by the OCMT procedure. Then, for all $\zeta > 0$, we have $E|FPR_{n,T}| = o(n^{\zeta-1}) + O\left[\exp(-n^{C_0})\right]$, for some finite positive constant C_0, where $FPR_{n,T}$ is defined by (20). If condition (5) holds instead of condition (4), then, assuming $\gamma > \frac{1}{2}\kappa_1^2$, we have $FPR_{n,T} \to_p 0$.*

4.2 Dynamic Extensions

An important assumption made so far is that noise variables are martingale difference processes which is restrictive in the case of time series applications. This assumption can be relaxed. In particular, under the less restrictive assumption that noise variables are exponentially mixing, it can be shown that all the theoretical results derived above hold. Details are provided in Section C of the online theory supplement. A further extension involves relaxing the martingale difference assumption for the signals and pseudo-signals. If we are willing to assume that either u_t is normally distributed or the covariates are deterministic, then a number of results become available. The relevant lemmas for the deterministic case are presented in Section

E of the online supplement. Alternatively, signals and pseudo-signals can be assumed to be exponentially mixing. In this general case, similar results to those in Theorems 1 and 2 can still be obtained. These are described in Section C of the online supplement. In the light of these theoretical extensions, one can also allow the DGP, (6), to include lagged dependent variables, $\mathbf{y}_{t,h} = (y_{t-1}, y_{t-2}, ..., y_{t-h})'$, where h is unknown. The OCMT procedure can now be applied to $\mathbf{x}_t$ augmented with $\mathbf{y}_{t,h_{\max}}$, where $h_{\max}$ is a maximum lag order selected by the investigator.

5 A Monte Carlo Study

We employ five different Monte Carlo (MC) designs, with or without lagged values of y_t. We allow the covariates to be serially correlated and consider different degrees of correlations across them. In addition, we experiment with Gaussian and non-Gaussian errors.

5.1 Data-generating processes (DGPs)

5.1.1 Design I (no hidden signals and no pseudo-signals)

y_t is generated as:

$$y_t = \varphi y_{t-1} + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \varsigma u_t, \tag{39}$$

where $u_t \sim IIDN(0,1)$ in the Gaussian case, and $u_t = [\chi^2_t(2) - 2]/2$ in the non-Gaussian case, in which $\chi^2_t(2)$ are independent draws from a χ^2-distribution with 2 degrees of freedom, for $t = 1, 2, ..., T$. We consider the 'static' specification with $\varphi = 0$, and two 'dynamic' specifications with $\varphi = 0.4$ and 0.8.[7] We set $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ and consider the following alternative ways of generating $\boldsymbol{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})'$:

DGP-I(a) Temporally uncorrelated and weakly collinear covariates: Signal variables are generated as $x_{it} = (\varepsilon_{it} + \nu g_t)/\sqrt{1+\nu^2}$, for $i = 1, 2, 3, 4$, and noise variables are generated as $x_{5t} = \varepsilon_{5t}$, $x_{it} = (\varepsilon_{i-1,t} + \varepsilon_{it})/\sqrt{2}$, for $i > 5$, where g_t and ε_{it} are independent draws either from $N(0,1)$ or from $[\chi^2_t(2) - 2]/2$, for $t = 1, 2, ..., T$, and $i = 1, 2, ..., n$. We set $\nu = 1$, which implies 50% pair-wise correlation among the signal variables.

DGP-I(b) Temporally correlated and weakly collinear covariates: Covariates are generated as in DGP-I(a), but with $\varepsilon_{it} = \rho_i \varepsilon_{i,t-1} + \sqrt{1-\rho_i^2} e_{it}$, in which $e_{it} \sim IIDN(0,1)$ or $IID[\chi^2_t(2) - 2]/2$. We set $\rho_i = 0.5$ for all i.

DGP-I(c) Strongly collinear noise variables due to a persistent unobserved common factor: Signal variables are generated as $x_{it} = (\varepsilon_{it} + g_t)/\sqrt{2}$, for $i = 1, 2, 3, 4$, and noise variables are generated as $x_{5t} = (\varepsilon_{5t} + b_i f_t)/\sqrt{3}$ and $x_{it} = \left[(\varepsilon_{i-1,t} + \varepsilon_{it})/\sqrt{2} + b_i f_t\right]/\sqrt{3}$, for $i > 5$, where $b_i \sim IIDN(1,1)$, $f_t = 0.95 f_{t-1} + \sqrt{1 - 0.95^2} v_t$, and v_t, g_t and ε_{it} are independent draws from $N(0,1)$ or $[\chi^2_t(2) - 2]/2$.

[7] Dynamic processes are initialized from zero starting values and the first 100 observations are discarded.

DGP-I(d) Low or high pair-wise correlation of signal variables: Covariates are generated as in DGP-I(a), but we set $\nu = \sqrt{\omega/(1-\omega)}$, for $\omega = 0.2$ (low pair-wise correlation) and 0.8 (high pair-wise correlation). This ensures that average correlation among the signals is ω.

5.1.2 Design II (featuring pseudo-signals)

The DGP is given by (39) and $\boldsymbol{x}_{nt}$ is generated as:

DGP-II(a) Two pseudo-signals: Signal variables are generated as $x_{it} = (\varepsilon_{it} + g_t)/\sqrt{2}$, for $i = 1, 2, 3, 4$, pseudo-signal variables are generated as $x_{5t} = \varepsilon_{5t} + \kappa x_{1t}$, and $x_{6t} = \varepsilon_{6t} + \kappa x_{2t}$, and noise variables are generated as $x_{it} = (\varepsilon_{i-1,t} + \varepsilon_{it})/\sqrt{2}$, for $i > 6$, where, as before, g_t, and ε_{it} are independent draws from $N(0,1)$ or $[\chi^2_t(2) - 2]/2$. We set $\kappa = 1.33$ (to achieve 80% correlation between the signal and the pseudo-signal variables).
DGP-II(b) All variables collinear with signals: $\boldsymbol{x}_{nt} \sim IID(\mathbf{0}, \boldsymbol{\Sigma}_x)$ with the elements of $\boldsymbol{\Sigma}_x$ given by $0.5^{|i-j|}$, $1 \leq i, j \leq n$. We generate $\boldsymbol{x}_{nt}$ with Gaussian and non-Gaussian innovations. In particular, $\boldsymbol{x}_{nt} = \boldsymbol{\Sigma}_x^{1/2}\boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \varepsilon_{2t}, ..., \varepsilon_{nt})'$, and ε_{it} are generated as independent draws from $N(0,1)$ or $[\chi^2_t(2) - 2]/2$.

5.1.3 Design III (featuring hidden signals)

y_t is generated by (39), $\boldsymbol{x}_{nt}$ is generated as in DGP-I(a), and the slope coefficients for the signals in (39) are selected so that, conditional on y_{t-1}, $\theta_4 = 0$:

DGP-III The fourth variable is hidden signal: We set $\beta_1 = \beta_2 = \beta_3 = 1$ and $\beta_4 = -1.5$. This implies $\theta_i \neq 0$ for $i = 1, 2, 3$ and $\theta_i = 0$ for $i \geq 4$, conditional on y_{t-1}.

5.1.4 Design IV (featuring both hidden signals and pseudo-signals)

In this case y_t is generated by (39), and:

DGP-IV(a) We generate $\boldsymbol{x}_{nt}$ in the same way as in DGP-II(a) which features two pseudo-signal variables. We generate slope coefficients β_i as in DGP-III to ensure $\theta_i \neq 0$ for $i = 1, 2, 3$, and $\theta_i = 0$ for $i = 4$, conditional on y_{t-1}.
DGP-IV(b) We generate $\boldsymbol{x}_{nt}$ in the same way as in DGP-II(b), where all covariates are collinear with signals. We set $\beta_1 = -0.875$ and $\beta_2 = \beta_3 = \beta_4 = 1$. This implies $\theta_i = 0$ for $i = 1$ and $\theta_i > 0$ for all $i > 1$, conditional on y_{t-1}.

5.1.5 Design V (Many signals)

For this design the DGP (**DGP-V**) is given by

$$y_t = \varphi y_{t-1} + \textstyle\sum_{i=1}^{n} i^{-2} x_{it} + \varsigma u_t, \tag{40}$$

where $\boldsymbol{x}_{nt}$ are generated as in design DGP-II(b), and u_t is generated in the same way as before. This design is inspired by the literature on approximately sparse models (Belloni et al. (2014b)).

Autoregressive processes are generated with zero starting values and 100 burn-in periods. ς is set so that $R^2 = 30\%$, 50% or 70% (on average) in static specifications ($\varphi = 0$). We do not change any parameters of the designs with an increase in φ, and we refer to the three R^2 measures corresponding to the three choices of ς as a low, medium and high fit. The sample combinations, $n = (100, 200, 300)$ and $T = (100, 300, 500)$ are considered, and all experiments are carried out using $R_{MC} = 2,000$ replications.

5.2 Variable selection methods

We consider six variable selection procedures, namely OCMT, Lasso, Adaptive Lasso (A-Lasso), Hard thresholding, SICA, and Boosting. In static specifications, the OCMT method is implemented as outlined in Section 3, where $c_p(n, \delta)$ is defined by (15) with $f(n, \delta) = n^{\delta}$ in the first stage and $f(n, \delta^*) = n^{\delta^*}$ in the subsequent stages. We use $p = 0.01$, and in line with the theoretical derivations we set $\delta = 1$ and $\delta^* = 2$. An online MC supplement provides results for other choices of $p \in \{0.01, 0.05, 0.1\}$ and $(\delta, \delta^*) \in \{(1, 1.5), (1, 2)\}$. It turns out that the choice of p is of second order importance. In the dynamic case, we augment the set of n covariates with $h_{\max} = 4$ lags of the dependent variable. Penalised regressions are implemented using the same set of possible values for the penalisation parameter λ as in Zheng et al. (2014), and following the literature λ is selected using 10-fold cross-validation. All methods are described in detail in the online MC supplement.

5.3 Monte Carlo results

We begin by reporting on the number of stages, denoted by $\hat{P}$, taken by OCMT before completion. This is important since our theory suggests that it should be close to P_0, which is 1 for DGPs I, II, and V without hidden signals, and 2 in the case of DGPs III and IV that do contain hidden signals. Realizations of $\hat{P}$ are very close to P_0 for both groups of experiments. The average number of stages in the two groups of experiments is $\overline{\hat{P}} = 1.03$ and 1.78, respectively. In addition, the frequency of MC replications with $\hat{P} > P_0$ and $\hat{P} > P_0 + 1$ turn out to be very small and amounted to 1.6%, and 0.003%, respectively.

Next, we focus on the average performance of Lasso, adaptive Lasso and OCMT methods, whilst the full set of results for all experiments and all six variable selection procedures is given in the online supplement. In our comparisons we focus on Lasso and adaptive Lasso since these are the main penalised regression methods used in the literature and also because they tend to perform better than Boosting. In our evaluation we use the following criteria: the true positive rate (TPR) defined by (20), the false positive rate (FPR) defined by (20), the false discovery rate of the true model (FDR*) defined by (33), the false discovery of the approximating model (FDR) defined by (21), the out-of-sample root mean square forecast error (RMSFE), and the

root mean square error of $\tilde{\boldsymbol{\beta}}$ (RMSE$_{\tilde{\boldsymbol{\beta}}}$).[8] We find that no method uniformly outperforms in the set of experiments we consider. This is true for the full set of methods (OCMT, Lasso, adaptive Lasso, Hard thresholding, SICA and Boosting) reported in the online supplement. The performance of individual methods can be quite different for individual experiments, and a relative assessment of these methods is provided in Table 1, which reports the fraction of experiments (in percent) where OCMT is outperformed by Lasso and Adaptive Lasso. These results clearly show that no method universally dominates. But it is interesting that the fraction of such experiments where OCMT is beaten by its competitors is relatively small, at most 22% for RMSFE and RMSE$_{\tilde{\boldsymbol{\beta}}}$ entries, in all experiments with the exception of dynamic specifications with $\varphi = 0.8$.

Summary statistics across the three choices of R^2 (low medium and high) and all the sample sizes ($n = 100, 200, 300$ and $T = 100, 300, 500$), for each of the five DGPs and with or without the lagged dependent variable, are reported Table A.1 in the Appendix. Lasso's TPR is in the majority of experiments larger than OCMT's, but so is the FPR and FDR as Lasso tends to overestimate the number of signals, which is well known in the literature. Adaptive Lasso in turn achieves better FPR and FDR outcomes compared with Lasso, but the performance of adaptive Lasso can be worse for TPR, RMSFE and RMSE$_{\tilde{\boldsymbol{\beta}}}$ in these experiments. The reported RMSFE and RMSE$_{\tilde{\boldsymbol{\beta}}}$ averages of Lasso and Adaptive Lasso are outperformed by OCMT in static specifications and dynamic specifications with low value of $\varphi = 0.4$ in Table A.1, by about 1.6% to 3.4%, and 9.1% to 40%, respectively. OCMT is very successful at eliminating the noise variables. On the other hand, the power of OCMT procedure to pick up the signals rises with $\sqrt{T}\left|\theta_{i,(j)}\right| / \sigma_{e_i,(T)}\sigma_{x_i,(T)}$, see Remark 1.[9] Hence the magnitude of $\theta_{i,(j)}$, T and R^2 are all important for the power of the OCMT. For instance, detailed findings reported in the online supplement show that an increase in the collinearity among signal variables, which results in a larger $\theta_{i,(j)}$, improves the performance of OCMT, but it worsens the performance of Lasso, since a higher collinearity of signal variables diminishes the marginal contribution of signals to the fit of the model. The performance of OCMT method also deteriorates with an increase in φ, and we see that in dynamic specifications with $\varphi = 0.8$ reported in the bottom panel of Table A.1, OCMT is beaten by Lasso and/or Adaptive Lasso in some instances. Findings for the non-Gaussian experiments are presented in Table A.2 in Appendix, which shows that the effects of allowing for non-Gaussian innovations seem to be rather marginal.

Overall, the small sample evidence suggests that the OCMT method is a valuable alternative to penalised regressions, since, in many cases, it can outperform the penalised regressions, that have become the *de facto* benchmark in the literature.

[8] RMSE$_{\tilde{\boldsymbol{\beta}}}$ is the square root of the trace of the MSE matrix of $\tilde{\boldsymbol{\beta}}$. Additional summary statistics, including the frequency of selecting the true model, and the statistics summarizing the distribution of the number of selected covariates are reported in the online supplement.

[9] $\sigma_{e_i,(T)}$ and $\sigma_{x_i,(T)}$ are defined by (B.49) in the online theory supplement, replacing $\mathbf{e}$, $\boldsymbol{x}$, and $\boldsymbol{M}_q$ by $\mathbf{e}_i$, $\boldsymbol{x}_i$, and $\boldsymbol{M}_{(j-1)}$, respectively.

6 Empirical Illustration

In this section we present an empirical application that highlights the utility of OCMT. In particular, we present a macroeconomic forecasting exercise for US GDP growth and CPI inflation using a large set of macroeconomic variables. The data set is quarterly and comes from Stock and Watson (2012). We use the smaller data set considered in Stock and Watson (2012), which contains 109 series. The series are transformed by taking logarithms and/or differencing following Stock and Watson (2012).[10] The transformed series span 1960Q3 to 2008Q4 and are collected in the vector $\boldsymbol{\xi}_t$ together with the target variable y_t (either US GDP growth or differenced log CPI inflation). Our estimation period is from 1960Q3 to 1990Q2 (120 periods) while the forecast evaluation period is 1990Q3 to 2008Q4. We produce one step ahead forecasts using five different procedures:[11] (a) AR benchmark with the number of lags selected by Schwarz Bayesian criterion (SBC) with maximum lag set equal to $h_{\max}$; (AR), (b) AR augmented with one lag of principal components, and the number of lags of the dependent variable is selected by SBC with maximum lag $h_{\max}$; (factor-augmented AR), (c-d) Lasso and adaptive Lasso regressions of the target variable y_t on lagged principal components, $\boldsymbol{\xi}_{t-1}$, and $h_{\max}$ lags of y_t. For Lasso and adaptive Lasso regressions, both the target variable and regressors are demeaned, and the regressors are normalised to have unit variances. (e) OCMT procedure is applied to regressions of y_t conditional on lagged principal components (included as pre-selected regressors), with $\boldsymbol{\xi}_{t-1}$ and $h_{\max}$ lags of y_t considered for variable selection. We set $\delta = 1$ in the first stage of OCMT, and $\delta^* = 2$ in the subsequent stages. We consider $p = 0.05$ below and findings for $p = 0.01$ and 0.1 are reported in the online empirical supplement. In all three data-rich procedures (b) to (e), the principal components are selected in a rolling scheme by the PC_{p_1} Bai and Ng (2002) criterion (with the maximum number of PCs set to 5). The maximum number of lags for the dependent variable, $h_{\max}$, is set to 4. We generate rolling forecasts using a rolling window of 120 observations.

We evaluate the forecasting performance of the methods using relative RMSFE where the AR forecast is the benchmark. Relative RMSFE statistics for the whole evaluation sample as well as for the pre-crisis sub-period (1990Q3-2007Q2) are reported in Table 2. In the case of GDP growth forecasts, we note that factor-augmented AR, Lasso and OCMT methods perform better than the AR benchmark. OCMT performs the best while Adaptive Lasso is the worst performer. However, the performance of the best methods is very close.[12] The differences in RMSFE in the case of inflation, reported in the bottom half of Table 2, are also relatively small with the factor-augmented AR(1) performing the best followed by OCMT and Lasso.

Variable inclusion frequencies are reported in Table 3, using the full evaluation sample.

[10] For further details, see the online supplement of Stock and Watson (2012), in particular columns E and T of their Table B.1.

[11] Further detail is provided in the online empirical supplement.

[12] Diebold-Mariano test statistics for all pairwise method comparisons can be found in the online supplement. The RMSFE differences among the best performing methods are not generally statistically significant.

Interestingly, for forecasting growth, the first lag of the dependent variable is among the most selected variables using OCMT (with the inclusion frequency of 45.9%), while no lags of the dependent variable are selected in the case of Lasso in any of the rolling windows. Results are different when inflation is considered. In this case, the inclusion frequency of the first lag of the dependent variable is 100% for both OCMT and Lasso methods. OCMT selects considerably fewer number of variables as compared to Lasso, an outcome that mirrors the Monte Carlo findings. In summary, we see that there is no method that uniformly outperforms all competitor methods and that OCMT is not far behind the best performing method.

7 Conclusion

Model selection is a recurring and fundamental topic in econometric analysis. This problem has become considerably more difficult for large-dimensional data sets where the set of possible specifications rise exponentially with the number of available covariates. In the context of linear regression models, penalised regression has become the *de facto* benchmark method of choice. However, issues such as the choice of penalty function and tuning parameters remains contentious.

In this paper, we provide an alternative approach based on multiple testing that is computationally simple, fast, and effective for sparse regression functions. Extensive theoretical and Monte Carlo results highlight these properties. In particular, we find that although no single method dominates across the broad set of experiments we considered, our proposed method can in many instances outperform existing penalised regression methods, whilst at the same time being computationally much faster by some orders of magnitude.

There are a number of avenues for future research. We have already considered the possibility of allowing for dynamics, but further extensions to more general settings with weakly exogenous regressors is clearly desirable. For empirical economic applications it is also important to allow for the possibility of weak and strong common factors affecting both the signal and pseudo-signal variables. A further possibility is to extend the idea of considering regressors individually to other testing frameworks, such as tests of forecasting ability. It is hoped that the results presented in this paper provide a basis for such further developments and empirical applications.

Table 1: Fraction of experiments (in percent) where OCMT is beaten by Lasso (L) and Adpaptive Lasso (A-L)

	Experiments with Gaussian innovations										Experiments with non-Gaussian innovations									
DGP type:	**DGP-I**		**DGP-II**		**DGP-III**		**DGP-IV**		**DGP-V**		**DGP-I**		**DGP-II**		**DGP-III**		**DGP-IV**		**DGP-V**	
No. of experiments:	135		54		27		54		27		135		54		27		54		27	
OCMT beaten by(*):	L	A-L	L	A-L	L	A-L	L	A-L	L	A-L	L	A-L	L	A-L	L	A-L	L	A-L	L	A-L
Static Specifications																				
TPR	15.6	6.7	20.4	3.7	44.4	29.6	59.3	38.9	100.0	3.7	17.8	6.7	22.2	3.7	48.1	22.2	66.7	38.9	96.3	3.7
FPR	0.0	0.0	0.0	18.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.2	0.0	0.0	0.0	0.0	0.0	0.0
FDR* (true model)	0.0	0.0	0.0	46.3	0.0	0.0	0.0	13.0	0.0	0.0	0.0	0.0	0.0	48.1	0.0	0.0	0.0	16.7	0.0	0.0
FDR (approximating model)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RMSFE	2.2	0.7	0.0	0.0	11.1	3.7	5.6	1.9	0.0	0.0	2.2	1.5	0.0	0.0	11.1	3.7	5.6	1.9	0.0	0.0
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	8.9	0.7	14.8	0.0	11.1	3.7	5.6	1.9	0.0	0.0	14.8	0.7	14.8	1.9	11.1	3.7	9.3	1.9	0.0	0.0
Dynamic Specifications																				
	Experiments with $\varphi = 0.4$																			
TPR	30.4	13.3	38.9	16.7	55.6	40.7	64.8	51.9	100.0	44.4	33.3	16.3	38.9	18.5	55.6	44.4	70.4	57.4	100.0	33.3
FPR	0.0	0.0	0.0	9.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	13.0	0.0	0.0	0.0	0.0	0.0	0.0
FDR* (true model)	0.0	1.5	0.0	33.3	0.0	0.0	0.0	9.3	0.0	0.0	0.0	1.5	0.0	35.2	0.0	0.0	0.0	11.1	0.0	0.0
FDR (approximating model)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RMSFE	8.9	7.4	11.1	7.4	22.2	18.5	16.7	13.0	3.7	3.7	9.6	9.6	11.1	9.3	22.2	22.2	16.7	18.5	11.1	11.1
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	14.8	2.2	11.1	0.0	11.1	11.1	11.1	5.6	0.0	0.0	14.8	2.2	13.0	0.0	11.1	14.8	14.8	7.4	0.0	0.0
	Experiments with $\varphi = 0.8$																			
TPR	64.4	43.0	75.9	61.1	66.7	66.7	83.3	83.3	100.0	100.0	71.9	42.2	81.5	61.1	70.4	66.7	85.2	83.3	100.0	100.0
FPR	20.0	65.9	0.0	53.7	0.0	14.8	0.0	27.8	0.0	70.4	20.0	69.6	0.0	59.3	0.0	14.8	0.0	27.8	0.0	74.1
FDR* (true model)	10.4	90.4	1.9	85.2	3.7	40.7	0.0	66.7	3.7	100.0	11.1	91.9	1.9	90.7	0.0	37.0	0.0	70.4	3.7	100.0
FDR (approximating model)	0.0	10.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RMSFE	24.4	42.2	3.7	24.1	14.8	29.6	14.8	40.7	0.0	7.4	25.9	41.5	7.4	14.8	14.8	22.2	14.8	38.9	0.0	0.0
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	60.0	45.2	55.6	44.4	44.4	37.0	55.6	51.9	40.7	0.0	60.7	44.4	55.6	37.0	44.4	44.4	55.6	57.4	29.6	0.0

Notes: (*) L: Lasso, A-L: Adaptive Lasso. DGPs I-IV are given by (39) and DGP V is given by (40). In the static case, DGP does not include lag dependent variable and selection of lags of the dependent variable is not considered. In the dynamic case, DGP includes one lag of the dependent variable, and the selection of up to $h_{\max} = 4$ lags of the dependent variable is considered. TPR (FPR) is the true (false) positive rate. FDR* is the false discovery rate for the true model and FDR is the false discovery rate for the approximating model. RMSFE is the root mean square forecast error. $\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$ is the root mean square error of $\tilde{\boldsymbol{\beta}}$. In DGP V, TPR is computed assuming that covariates $i = 1, 2, ..., 11$ are the signal variables, and FPR and FDR are computed assuming covariates $i > 11$ are the noise variables. In the case of Oracle method the identity of rue variables is known. In DGP V, Oracle* method assumes the first 11 covariates are the signal variables. Lasso is implemented using the same set of possible values for the penalisation parameter λ as in Zheng et al. (2014), and λ is selected using 10-fold cross-validation. Adaptive Lasso method is implemented as described in Section 2.8.4 of Buhlmann and van de Geer (2011) based on the implementation of the Lasso method described above. OCMT results are based on $p = 0.01$, $\delta = 1$ in the first stage, and $\delta^* = 2$ in the subsequent stages of the OCMT procedure. See Section 5 for further details. The complete set of findings is reported in the online MC supplement.

Table 2: RMSFE performance of the AR, factor-augmented AR, Lasso and OCMT methods

Evaluation sample:	Full 1990Q3-2008Q4		Pre-crisis 1990Q3-2007Q2	
	RMSFE (×100)	Relative RMSFE	RMSFE (×100)	Relative RMSFE
	Real output growth			
AR benchmark	0.561	1.000	0.505	1.000
Factor-augmented AR	0.484	0.862	0.470	0.930
Lasso	0.510	0.910	0.465	0.922
Adaptive Lasso	0.561	1.000	0.503	0.996
OCMT	0.477	0.850	0.461	0.912
	Inflation			
AR benchmark	0.601	1.000	0.435	1.000
Factor-augmented AR	0.557	0.927	0.415	0.954
Lasso	0.599	0.997	0.462	1.063
Adaptive Lasso	0.715	1.190	0.524	1.205
OCMT	0.590	0.982	0.464	1.068

Notes: RMSFE is computed based on rolling forecasts with a rolling window of 120 observations. The source of the data is the smaller data set with 109 time series provided by Stock and Watson (2012). The series are transformed by taking logarithms and/or differencing following Stock and Watson (2012). The transformed series span 1960Q3 to 2008Q4 and are collected in the vector $\boldsymbol{\xi}_t$. Set of regressors in Lasso and adaptive-Lasso contains $h_{\max} = 4$ lags of y_t (lagged target variables), $\boldsymbol{\xi}_{t-1}$, and a lagged set of principal components obtained from the large data set given by $(y_t, \boldsymbol{\xi}_t')'$. OCMT procedure is applied to regressions of y_t conditional on lagged principal components (included as pre-selected regressors) with $\boldsymbol{\xi}_{t-1}$ and $h_{\max} = 4$ lags of y_t considered for variable selection. OCMT is reported for $p = 0.05$ and $\delta = 1$ in the first stage, and $p = 0.05$ and $\delta^* = 2$ in the subsequent stages of the OCMT procedure. The number of principal components in the factor-augmented AR, Lasso, adaptive-Lasso, and OCMT methods is determined in a rolling scheme by using criterion PC_{p_1} of Bai and Ng (2002) (with the maximum number of PCs set to 5). See Section 6 and the online empirical supplement for further details.

Table 3: Top 5 variables with highest inclusion frequencies based on the Lasso and OCMT selection methods

Output growth			
Lasso		**OCMT**	
1. Real gross private domestic investment - residential (*)	100.0%	1. Residential price index	47.3%
2. Real personal consumption expenditures - services (*)	100.0%	2. First lag of the dependent variable	45.9%
3. Employees, nonfarm - mining	89.2%	3. Industrial production index - fuels	43.2%
4. Index of help - wanted advertising in newspapers	75.7%	4. Labor productivity (output per hour)	37.8%
5. Employment: Ratio; Help-wanted ads: No. unemployed CLF	56.8%	5. Employees, nonfarm - mining	27.0%
Average number of selected variables	8.1	Average number of selected variables (excluding pre-selected factors)	2.2
Inflation			
Lasso		**OCMT**	
1. Interest rate: U.S. Treasury bills, sec. mkt, 3-mo (% per ann)	100.0%	1. First lag of the dependent variable	100.0%
2. Real personal consumption expenditures - services (*)	100.0%	2. Third lag of the dependent variable	78.4%
3. First lag of the dependent variable	100.0%	3. MZM money stock (FRB St. Lois)	71.6%
4. Employees, nonfarm - mining	98.6%	4. Money stock: M2	45.9%
5. Second lag of the dependent variable	98.6%	5. Recreation price index	33.8%
Average number of selected variables	21.7	Average number of selected variables (excluding pre-selected factors)	4.0

Notes: This table reports the top 5 highest inclusion frequencies of the variables selected using the Lasso and OCMT procedure on the full evaluation sample, 1990Q3-2008Q4. OCMT is reported $p = 0.05$ and for $\delta = 1$ in the first stage, and $\delta^* = 2$ in the subsequent stages of the OCMT procedure.
(*) quantity index.

A Appendix

A.1 Additional notations and definitions

Throughout this appendix we consider the following events:

$$\mathcal{A}_0 = \mathcal{H} \cap \mathcal{G}, \text{ where } \mathcal{H} = \{\textstyle\sum_{i=1}^{k} \widehat{\mathcal{J}}_i = k\}, \text{ and } \mathcal{G} = \{\textstyle\sum_{i=k+k^*+1}^{n} \widehat{\mathcal{J}}_i = 0\}. \tag{A.1}$$

$\mathcal{A}_0$, also defined by (28), is the event of selecting the approximating model, $\mathcal{H}$ is the event that all signals are selected, and $\mathcal{G}$ is the event that no noise variable is selected. We also denote the event that exactly j noise variables are selected by $\mathcal{G}_j = \{\sum_{i=k+k^*+1}^{n} \widehat{\mathcal{J}}_i = j\}$, for $j = 0, 1, ..., n-k-k^*$, with $\mathcal{G} \equiv \mathcal{G}_0$. For the analysis of different stages of OCMT, we also introduce the event $\mathcal{B}_{i,s}$, which is the event that variable i is selected at the s^{th} stage of the OCMT procedure. $\mathcal{L}_{i,s} = \cup_{h=1}^{s} \mathcal{B}_{i,h}$ is the event that variable i is selected up to and including stage s, namely in any of the stages $j = 1, 2, ..., s$ of the OCMT procedure, and $\mathcal{L}_s = \cap_{i=1}^{k} \mathcal{L}_{i,s}$ is the event that all signals are selected up to and including stage s of the OCMT procedure. $\mathcal{T}_s$ is the event that OCMT stops after s stages or less. $\mathcal{D}_{s,T}$ is the event that the number of variables selected in the first s stages of OCMT ($\hat{k}_{(j)}$, $j = 1, 2, ..., s$) is smaller than or equal to l_T, where $l_T = \ominus(n^{\nu})$ and ν satisfies $\epsilon < \nu < \kappa_1/3$. Note that when $T = \ominus(n^{\kappa_1})$ then $l_T = \ominus\left(T^{\nu/\kappa_1}\right) = o\left(T^{1/3}\right)$ for $\nu < \kappa_1/3$.

Notations: Let $\mathbf{a} = (a_1, a_2, ..., a_n)'$ and $\mathbf{A} = (a_{ij})$ be an $n \times 1$ vector and an $n \times m$ matrix, respectively. Then, $\|\mathbf{a}\| = (\Sigma_{i=1}^{n} a_i^2)^{1/2}$ and $\|\mathbf{a}\|_1 = \Sigma_{i=1}^{n} |a_i|$ are the Euclidean (L_2) and L_1 norms of $\mathbf{a}$, respectively. $\|\boldsymbol{A}\|_F = [Tr(\boldsymbol{A}\boldsymbol{A}')]^{1/2}$ is the Frobenius norm of $\mathbf{A}$.

A.2 Proofs of Propositions and Theorems

All proofs are based on the set of lemmas presented and established in the online theory supplement. In particular, Lemmas A1-A9 are auxiliary ones, mostly providing supporting results for the main lemma of the paper, namely Lemma A10, which provides the basic exponential inequalities that underlie most of our results. A simple version of this lemma is included in the paper as Proposition 2.

A.2.1 Proof of Proposition 1

We recall that P_0 is a population quantity. This formally means that, to determine P_0, OCMT is carried out assuming $\Pr[|t_{\hat{\phi}_{i,(j)}}| > c_p(n,\delta) \,|\theta_{i,(j)} \neq 0] = 1$, and $\Pr[|t_{\hat{\phi}_{i,(j)}}| > c_p(n,\delta) \,|\theta_{i,(j)} = 0] = 0$ for all i, j. So, if $\theta_{i,(1)} \neq 0$, for all i for which $\beta_i \neq 0$, it obviously follows that $P_0 = 1$. Next, assume that the subset of signal variables in $\mathbf{X}_k$, such that for each element of this subset, $\theta_{i,(1)} = 0$, is not empty. Then, these signals will not be selected in the first stage of OCMT. By Lemma A1 in the online supplement, it follows that the subset of signals for which $\theta_{i,(1)} = 0$ is smaller than the set of signals and therefore at least one signal will be picked up in the first

stage of OCMT. It then follows, by Lemma A1, that in the second stage of OCMT, at least one hidden signal, for which $\theta_{i,(1)} = 0$ will have $\theta_{i,(2)} \neq 0$. Therefore, such hidden signal(s) will be picked up in the second stage. Proceeding recursively using Lemma A1, it then follows that all hidden signals for which $\theta_{i,(1)} = 0$, will satisfy $\theta_{i,(j)} \neq 0$ for some $j \leq k$, proving the proposition.[13]

A.2.2 Proof of Theorem 1

Noting that $\mathcal{T}_k$ is the event that the OCMT procedure stops after k stages or less, we have $\Pr\left(\hat{P} > k\right) = \Pr\left(\mathcal{T}_k^c\right) = 1 - \Pr\left(\mathcal{T}_k\right)$, where $\hat{P}$ is defined by (17). Substituting (B.83) of Lemma A20 in the online supplement for $\Pr\left(\mathcal{T}_k\right)$, we obtain, $\Pr\left(\hat{P} > k\right) = O\left(n^{1-\nu-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0 n^{C_1\kappa_1}\right)\right]$, for some $C_0, C_1 > 0$, any $\varkappa$ in $0 < \varkappa < 1$, and any ν in $0 \leq \epsilon < \nu < \kappa_1/3$, where $\kappa_1 > 0$ defines the rate for $T = \ominus\left(n^{\kappa_1}\right)$, and ϵ in $0 \leq \epsilon < \min\left\{1, \kappa_1/3\right\}$ defines the rate for $k^* = \ominus\left(n^{\epsilon}\right)$. But note that $O\left(n^{1-\nu-\varkappa\delta}\right)$ can be written equivalently as $O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)$. This follows since $1 - \kappa_1/3 - \varkappa\delta = 1 - \left(\kappa_1/3 - \varepsilon\delta\right) - \left(\varkappa + \varepsilon\right)\delta = 1 - \tilde{\nu} - \tilde{\varkappa}\delta$, where $\tilde{\nu} = \kappa_1/3 - \varepsilon\delta$ and $\tilde{\varkappa} = \varkappa + \varepsilon$, for $\varepsilon > 0$ sufficiently small. Specifically, setting $\varepsilon < \min\left\{1 - \varkappa, \left(\kappa_1/3 - \epsilon\right)/\delta\right\}$, it follows that $\tilde{\varkappa}$ and $\tilde{\nu}$ satisfy $0 < \tilde{\varkappa} < 1$ and $\epsilon < \tilde{\nu} < \kappa_1/3$, respectively, as required. Hence

$$\Pr(\hat{P} > k) = \Pr\left(\mathcal{T}_k^c\right) = O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0 n^{C_1\kappa_1}\right)\right], \tag{A.2}$$

for some $C_0, C_1 > 0$ and any $\varkappa$ in $0 < \varkappa < 1$. Noting that $O\left[n\exp\left(-C_0 n^{C_1\kappa_1}\right)\right] = O\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$ for any $0 < C_2 < C_1$, we have $\Pr\left(\hat{P} > k\right) = O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$, for some $C_2 > 0$, which establishes (29). Similarly, by (B.86) and noting that $n \geq n^{1-\nu}$ for $\nu \geq 0$, we also have (which is required subsequently)

$$\Pr\left(\mathcal{D}_{k,T}^c\right) = O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0 T^{C_1\kappa_1}\right)\right], \tag{A.3}$$

for some $C_0, C_1 > 0$ and any $\varkappa$ in $0 < \varkappa < 1$.

To establish result (30), we first note that

$$\Pr(\mathcal{A}_0^c) = \Pr(\mathcal{A}_0^c|\mathcal{D}_{k,T})\Pr(\mathcal{D}_{k,T}) + \Pr(\mathcal{A}_0^c|\mathcal{D}_{k,T}^c)\Pr(\mathcal{D}_{k,T}^c) \leq \Pr(\mathcal{A}_0^c|\mathcal{D}_{k,T}) + \Pr(\mathcal{D}_{k,T}^c), \tag{A.4}$$

where $\Pr(\mathcal{D}_{k,T}^c)$ is given by (A.3). Also using (A.1) we have $\mathcal{A}_0^c = \mathcal{H}^c \cup \mathcal{G}^c$, and hence

$$\Pr(\mathcal{A}_0^c|\mathcal{D}_{k,T}) \leq \Pr\left(\mathcal{H}^c|\,\mathcal{D}_{k,T}\right) + \Pr\left(\mathcal{G}^c|\,\mathcal{D}_{k,T}\right) = A_{n,T} + B_{n,T}, \tag{A.5}$$

where $\mathcal{H}$ and $\mathcal{G}$ are given by (A.1). Therefore $\mathcal{H}^c = \{\sum_{i=1}^k \widehat{\mathcal{J}}_i < k\}$, and $\mathcal{G}^c = \{\sum_{i=k+k^*+1}^n \widehat{\mathcal{J}}_i > 0\}$. Consider the terms $A_{n,T}$ and $B_{n,T}$, in turn:

$$A_{n,T} = \Pr\left(\mathcal{H}^c|\,\mathcal{D}_{k,T}\right) \leq \textstyle\sum_{i=1}^k \Pr(\widehat{\mathcal{J}}_i = 0\Big|\,\mathcal{D}_{k,T}). \tag{A.6}$$

[13] Note that this proposition allows the net effects to tend to zero with T (or n) at a sufficiently slow rate as set out in Assumption 6, as long as they are not exactly zero. See also Lemma A1 in the online supplement.

But, the event $\{\widehat{\mathcal{J}}_i = 0|\mathcal{D}_{k,T}\}$ can occur only if $\{\cap_{j=1}^{k}\mathcal{B}_{i,j}^{c}|\mathcal{D}_{k,T}\}$ occurs, while $\{\cap_{j=1}^{k}\mathcal{B}_{i,j}^{c}|\mathcal{D}_{k,T}\}$ can occur without $\{\widehat{\mathcal{J}}_i = 0|\mathcal{D}_{k,T}\}$ occurring. Therefore, $\Pr[\widehat{\mathcal{J}}_i = 0|\mathcal{D}_{k,T}] \leq \Pr(\cap_{j=1}^{k}\mathcal{B}_{i,j}^{c}|\mathcal{D}_{k,T})$. Then,

$$\begin{aligned}\Pr\left(\cap_{j=1}^{k}\mathcal{B}_{i,j}^{c}\middle|\,\mathcal{D}_{k,T}\right) = &\Pr\left(\mathcal{B}_{i,1}^{c}\middle|\,\mathcal{D}_{k,T}\right)\times\Pr\left(\mathcal{B}_{i,2}^{c}\middle|\,\mathcal{B}_{i,1}^{c},\mathcal{D}_{k,T}\right)\times\Pr\left(\mathcal{B}_{i,3}^{c}\middle|\,\mathcal{B}_{i,2}^{c}\cap\mathcal{B}_{i,1}^{c},\mathcal{D}_{k,T}\right)\\ &\times ...\times\Pr\left(\mathcal{B}_{i,k}^{c}\middle|\,\mathcal{B}_{i,k-1}^{c}\cap ...\cap\mathcal{B}_{i,1}^{c},\mathcal{D}_{k,T}\right). \qquad (A.7)\end{aligned}$$

But, by Proposition 1 we are guaranteed that for some j in $1 \leq j \leq k$, $\theta_{i,(j)} \neq 0$, $i = 1, 2, ..., k$. Therefore, for some j in $1 \leq j \leq k$,

$$\Pr(\mathcal{B}_{i,j}^{c}|\,\mathcal{B}_{i,j-1}^{c}\cap ...\cap\mathcal{B}_{i,1}^{c},\mathcal{D}_{k,T}) = \Pr(\mathcal{B}_{i,j}^{c}|\,\mathcal{B}_{i,j-1}^{c}\cap ...\cap\mathcal{B}_{i,1}^{c},\theta_{i,(j)}\neq 0,\mathcal{D}_{k,T}),$$

and by (B.52) of Lemma A10 in the online supplement, $\Pr(\mathcal{B}_{i,j}^{c}|\,\mathcal{B}_{i,j-1}^{c}\cap ...\cap\mathcal{B}_{i,1}^{c},\theta_{i,(j)}\neq 0,\mathcal{D}_{k,T}) = O\left[\exp\left(-C_0T^{C_1}\right)\right]$, for $i = 1, 2, ..., k$, and some $C_0, C_1 > 0$. Therefore,

$$\Pr(\widehat{\mathcal{J}}_i = 0 \mid \mathcal{D}_{k,T}) = O\left[\exp\left(-C_0T^{C_1}\right)\right], \text{ for } i = 1, 2, ..., k. \qquad (A.8)$$

Substituting this result in (A.6), we have

$$A_{n,T} = \Pr\left(\mathcal{H}^{c}|\,\mathcal{D}_{k,T}\right) \leq k\exp\left(-C_0T^{C_1}\right). \qquad (A.9)$$

Similarly, for $B_{n,T}$ we first note that

$$B_{n,T} = \Pr[\cup_{i=k+k^*+1}^{n}(\widehat{\mathcal{J}}_i > 0)|\mathcal{D}_{k,T}] \leq \textstyle\sum_{i=k+k^*+1}^{n}E(\widehat{\mathcal{J}}_i\,|\mathcal{D}_{k,T}). \qquad (A.10)$$

Also, $E(\widehat{\mathcal{J}}_i\,|\mathcal{D}_{k,T}) = E(\widehat{\mathcal{J}}_i\,|\mathcal{D}_{k,T},\mathcal{T}_k)\Pr(\mathcal{T}_k|\mathcal{D}_{k,T}) + E(\widehat{\mathcal{J}}_i\,|\mathcal{D}_{k,T},\mathcal{T}_k^{c})\Pr(\mathcal{T}_k^{c}|\mathcal{D}_{k,T}) \leq E(\widehat{\mathcal{J}}_i\,|\mathcal{D}_{k,T},\mathcal{T}_k) + \Pr(\mathcal{T}_k^{c}|\mathcal{D}_{k,T})$, since $E(\widehat{\mathcal{J}}_i\,|\mathcal{D}_{k,T},\mathcal{T}_k^{c}) \leq 1$. Hence $B_{n,T} \leq \sum_{i=k+k^*+1}^{n}E(\widehat{\mathcal{J}}_i\,|\mathcal{D}_{k,T},\mathcal{T}_k) + (n-k-k^*)\Pr(\mathcal{T}_k^{c}|\mathcal{D}_{k,T})$. Consider now the first term of the above and note that

$$\begin{aligned}\textstyle\sum_{i=k+k^*+1}^{n}E(\widehat{\mathcal{J}}_i\,|\mathcal{D}_{k,T},\mathcal{T}_k) = &\textstyle\sum_{i=k+k^*+1}^{n}\Pr[|t_{\hat{\phi}_{i,(1)}}| > c_p(n,\delta)\,|\theta_{i,(1)} = 0,\mathcal{D}_{k,T},\mathcal{T}_k]\\ &+\textstyle\sum_{i=k+k^*+1}^{n}\sum_{j=2}^{k}\Pr[|t_{\hat{\phi}_{i,(j)}}| > c_p(n,\delta^*)\,|\theta_{i,(j)} = 0,\mathcal{D}_{k,T},\mathcal{T}_k],\end{aligned}$$

where we have made use of the fact that the net effect coefficients, $\theta_{i,(j)}$, of noise variables are zero for $i = k+k^*+1, k+k^*+2, ..., n$ and all j. Also by (B.51) of Lemma A10 and result (ii) of Lemma A2, we have

$$\begin{aligned}&\sum_{i=k+k^*+1}^{n}\Pr\left(|t_{\hat{\phi}_{i,(1)}}| > c_p(n,\delta)\,|\theta_{i,(1)} = 0,\mathcal{D}_{k,T},\mathcal{T}_k\right) + \sum_{i=k+k^*+1}^{n}\sum_{s=2}^{k}\Pr\left(|t_{\hat{\phi}_{i,(s)}}| > c_p(n,\delta^*)\,|\theta_{i,(s)} = 0,\mathcal{D}_{k,T},\mathcal{T}_k\right)\\ &\leq (n-k-k^*)\exp\left[-\varkappa c_p^2(n,\delta)/2\right] + (k-1)(n-k-k^*)\exp\left[-\varkappa c_p^2(n,\delta^*)/2\right] + O\left[n\exp\left(-C_0T^{C_1}\right)\right]\\ &= O\left(n^{1-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0T^{C_1}\right)\right].\end{aligned}$$

Further, by (B.92), $n\Pr(\mathcal{T}_k^{c}|\mathcal{D}_{k,T}) = O\left(n^{2-\varkappa\delta^*}\right) + O\left[n^2\exp\left(-C_0T^{C_1}\right)\right]$, giving, overall,

$$B_{n,T} = O\left(n^{1-\delta\varkappa}\right) + O\left(n^{2-\delta^*\varkappa}\right) + O\left[n^2\exp\left(-C_0T^{C_1}\right)\right], \qquad (A.11)$$

where we used that $O\left[n\exp\left(-C_0T^{C_1}\right)\right]$ is dominated by $O\left[n^2\exp\left(-C_0T^{C_1}\right)\right]$, and $O\left(n^{1-\varkappa\delta^*}\right)$ is dominated by $O\left(n^{1-\varkappa\delta}\right)$ for $\delta^*>\delta>0$. Substituting for $A_{n,T}$ and $B_{n,T}$ from (A.9) and (A.11) in (A.5) and using (A.4) we obtain $\Pr(\mathcal{A}_0^c)\leq O\left(n^{1-\delta\varkappa}\right)+O\left(n^{2-\delta^*\varkappa}\right)+O\left[n^2\exp\left(-C_0T^{C_1}\right)\right]+\Pr(\mathcal{D}_{k,T}^c)$, where $\Pr(\mathcal{D}_{k,T}^c)$ is already given by (A.3), and $k\exp\left(-C_0T^{C_1}\right)$ is dominated by $O\left[n^2\exp\left(-C_0T^{C_1}\right)\right]$. Hence, noting that $\Pr\left(\mathcal{A}_0\right)=1-\Pr(\mathcal{A}_0^c)$, then

$$\Pr\left(\mathcal{A}_0\right)=1+O\left(n^{1-\delta\varkappa}\right)+O\left(n^{2-\delta^*\varkappa}\right)+O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)+O\left[n^2\exp\left(-C_0T^{C_1}\right)\right], \tag{A.12}$$

since $O[n\exp\left(-C_0T^{C_1}\right)]$ is dominated by $O[n^2\exp\left(-C_0T^{C_1}\right)]$, and $O(n^{1-\kappa_1/3-\varkappa\delta^*})$ is dominated by $O(n^{1-\kappa_1/3-\varkappa\delta})$, for $\delta^*>\delta>0$. Result (30) now follows noting that $T=\ominus\left(n^{\kappa_1}\right)$ and that $O\left[n^2\exp\left(-C_0n^{C_1\kappa_1}\right)\right]=O\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$ for some C_2 in $0<C_2<C_1$. If, in addition, $\delta>1$, and $\delta^*>2$, then $\Pr\left(\mathcal{A}_0\right)\to 1$, as n,$T\to\infty$, for any $\kappa_1>0$.

We establish result (32) next, before establishing results (31) and the result on FDR. Consider $FPR_{n,T}$ defined by (20), and note that the probability of noise or pseudo-signal variable i being selected in any stages of the OCMT procedure is given by $\Pr\left(\mathcal{L}_{i,n}\right)$, for $i=k+1,k+2,...,n$. Then

$$E\left|FPR_{n,T}\right|=\frac{\sum_{i=k+1}^{n}\Pr\left(\mathcal{L}_{i,n}\right)}{n-k}=\frac{\sum_{i=k+1}^{k+k^*}\Pr\left(\mathcal{L}_{i,n}\right)}{n-k}+\frac{\sum_{i=k+k^*+1}^{n}\Pr\left(\mathcal{L}_{i,n}\right)}{n-k}. \tag{A.13}$$

Since $\sum_{i=k+1}^{k+k^*}\Pr\left(\mathcal{L}_{i,n}\right)\leq k^*$ then

$$E\left|FPR_{n,T}\right|\leq(n-k)^{-1}k^*+(n-k)^{-1}\textstyle\sum_{i=k+k^*+1}^{n}\Pr\left(\mathcal{L}_{i,n}\right). \tag{A.14}$$

Note that

$$(n-k)^{-1}\textstyle\sum_{i=k+k^*+1}^{n}\Pr\left(\mathcal{L}_{i,n}\right)\leq(n-k)^{-1}\sum_{i=k+k^*+1}^{n}\Pr\left(\mathcal{L}_{i,n}|\mathcal{D}_{k,T}\right)+\Pr\left(\mathcal{D}_{k,T}^c\right). \tag{A.15}$$

Furthermore

$$\Pr\left(\mathcal{L}_{i,n}|\mathcal{D}_{k,T}\right)\leq\Pr\left(\mathcal{L}_{i,n}|\mathcal{D}_{k,T},\mathcal{T}_k\right)+\Pr\left(\mathcal{T}_k^c\right). \tag{A.16}$$

An upper bound to $\Pr\left(\mathcal{T}_k^c\right)=\Pr(\hat{P}>k)$ is established in the first part of this proof, see (A.2). We focus on $\Pr\left(\mathcal{L}_{i,n}|\mathcal{D}_{k,T},\mathcal{T}_k\right)$ next. Due to the conditioning on the event $\mathcal{T}_k$, we have $\Pr\left(\mathcal{L}_{i,n}|\mathcal{D}_{k,T},\mathcal{T}_k\right)=\Pr\left(\mathcal{L}_{i,k}|,\mathcal{D}_{k,T},\mathcal{T}_k\right)$, and in view of $\mathcal{L}_{i,k}=\cup_{h=1}^{k}\mathcal{B}_{i,h}$ we obtain

$$\Pr\left(\mathcal{L}_{i,k}|\mathcal{D}_{k,T},\mathcal{T}_k\right)\leq\textstyle\sum_{s=1}^{k}\Pr\left(\mathcal{B}_{i,s}|\theta_{i,(s)}=0,\mathcal{D}_{k,T},\mathcal{T}_k\right),\ \text{for } i>k+k^*, \tag{A.17}$$

where we note that $\Pr\left(\mathcal{B}_{i,s}|\mathcal{D}_{k,T},\mathcal{T}_k\right)=\Pr\left(\mathcal{B}_{i,s}|\theta_{i,(s)}=0,\mathcal{D}_{k,T},\mathcal{T}_k\right)$, for $i>k+k^*$ since the net effect coefficients of the noise variables at any stage of OCMT are zero. Further, using (B.51) of Lemma A10, for $i=k+k^*+1,k+k^*+2,...,n$, we have

$$\Pr\left(\mathcal{B}_{i,s}|\theta_{i,(s)}=0,\mathcal{D}_{k,T},\mathcal{T}_k\right)=\begin{cases}O\left\{\exp\left[-\varkappa c_p^2(n,\delta)/2\right]\right\}+O\left[\exp(-C_0T^{C_1})\right], & s=1\\ O\left\{\exp\left[-\varkappa c_p^2(n,\delta^*)/2\right]\right\}+O\left[\exp(-C_0T^{C_1})\right], & s>1\end{cases}, \tag{A.18}$$

where $\varkappa = [(1-\pi)/(1+d_T)]^2$. Clearly $0 < \varkappa < 1$, since $0 < \pi < 1$, and d_T is a bounded positive sequence. Hence, given result (ii) of Lemma A2 in the online supplement, for $i = k+k^*+1, k+k^*+2, ..., n$, we have

$$\sum_{s=1}^{k} \Pr\left(\mathcal{B}_{i,s}|\theta_{i,(s)}=0, \mathcal{D}_{k,T}, \mathcal{T}_k\right) = O\left(n^{-\delta\varkappa}\right) + O\left(n^{-\delta^*\varkappa}\right) + O\left[\exp(-C_0T^{C_1})\right].$$

Using this result in (A.17) and averaging across $i = k+k^*+1, k+k^*+2, ..., n$, we obtain

$$(n-k)^{-1}\sum_{i=k+k^*+1}^{n} \Pr\left(\mathcal{L}_{i,k}|\mathcal{D}_{k,T}, \mathcal{T}_k\right) = O\left(n^{-\varkappa\delta}\right) + O\left(n^{-\varkappa\delta^*}\right) + O\left[\exp(-C_0T^{C_1})\right]. \quad (A.19)$$

Overall, with $\delta^* > \delta$, $T = \ominus(n^{\kappa_1})$, $k^* = \ominus(n^{\epsilon})$, and using (A.2), (A.3), (A.14)-(A.16) and (A.19), we have $E|FPR_{n,T}| = k^*/(n-k) + O\left(n^{-\varkappa\delta}\right) + O\left(n^{-\varkappa\delta^*}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[\exp(-C_0n^{C_1\kappa_1})\right] + O\left(n^{\epsilon-1}\right) + O\left[n\exp\left(-C_0n^{C_1\kappa_1}\right)\right]$. But $O\left[\exp(-C_0n^{C_1\kappa_1})\right]$ and $O\left[n\exp\left(-C_0n^{C_1\kappa_1}\right)\right]$ are dominated by $\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$ for some $0 < C_2 < C_1$. In addition, since $\delta^* > \delta$ and $\varkappa$ is positive, the terms $O\left(n^{-\varkappa\delta^*}\right)$ and $O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right)$ are dominated by $O\left(n^{-\varkappa\delta}\right)$ and $O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)$, respectively. Hence, $E|FPR_{n,T}| = k^*/(n-k) + O\left(n^{-\varkappa\delta}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{\epsilon-1}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$, for some $C_2 > 0$, which completes the proof of (32).

To establish (31) we note from (20) that

$$E|TPR_{n,T}| = k^{-1}\sum_{i=1}^{k} \Pr[\widehat{\mathcal{J}}_i = 1]. \quad (A.20)$$

But $\Pr[\widehat{\mathcal{J}}_i = 1] = 1 - \Pr[\widehat{\mathcal{J}}_i = 0]$, and $\Pr[\widehat{\mathcal{J}}_i = 0] \leq \Pr[\widehat{\mathcal{J}}_i = 0|\mathcal{D}_{k,T}] + \Pr\left(\mathcal{D}_{k,T}^c\right)$. Using (A.8) and (A.3) and dropping the terms $O\left[\exp\left(-C_0T^{C_1}\right)\right]$ and $O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right)$ that are dominated by $O\left[n\exp\left(-C_0T^{C_1}\right)\right]$ and $O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)$, respectively (noting that $\delta^* > \delta > 0$) we obtain $\Pr[\widehat{\mathcal{J}}_i = 0] = O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[n\exp\left(-C_0T^{C_1}\right)\right]$, for $i = 1, 2, ..., k$. Hence, $\sum_{i=1}^{k}\Pr[\widehat{\mathcal{J}}_i = 1] = k + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[n\exp\left(-C_0T^{C_1}\right)\right]$, which, after substituting this expression in (A.20), and noting that $T = \ominus(n^{\kappa_1})$, and
$O\left[n\exp\left(-C_0n^{C_1\kappa_1}\right)\right] = O\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$, for some C_2 in $0 < C_2 < C_1$ yields

$$E|TPR_{n,T}| = 1 + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[\exp\left(-n^{C_2\kappa_1}\right)\right], \quad (A.21)$$

for some $C_2 > 0$, as required.

To establish the result on FDR, we first note that

$$FDR_{n,T} = \frac{\sum_{i=1}^{n} I\left(\widehat{\mathcal{J}}_i = 1, \text{ and } \beta_i = \theta_i = 0\right)}{(n-k)FPR_{n,T} + kTPR_{n,T} + 1}.$$

Consider the numerator first. Taking expectations $E\sum_{i=1}^{n} I[\widehat{\mathcal{J}}_i = 1, \text{ and } \beta_i = \theta_i = 0] = \sum_{i=k+k^*+1}^{n}\Pr\left(\mathcal{L}_{i,n}\right)$. Using (A.2),(A.3),(A.15), and (A.16), and noting $T = \ominus(n^{\kappa_1})$, we have

$$\begin{aligned}\sum_{i=k+k^*+1}^{n}\Pr\left(\mathcal{L}_{i,n}\right) = {} & O\left(n^{1-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left(n^{2-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{2-\kappa_1/3-\varkappa\delta^*}\right) \\ & + O\left(n^{2-\varkappa\delta^*}\right) + O\left[n\exp(-C_0n^{C_1\kappa_1})\right] + O\left[n^2\exp\left(-C_0n^{C_1\kappa_1}\right)\right], \quad (A.22)\end{aligned}$$

for some $C_0, C_1 > 0$. Hence, if $\delta > \max\{1, 2 - \kappa_1/3\}$, and $\delta^* > 2$, then $\sum_{i=k+k^*+1}^{n} \Pr(\mathcal{L}_{i,n}) \to 0$, and

$$\sum_{i=1}^{n} I[\widehat{\mathcal{J}}_i = 1, \text{ and } \beta_i = \theta_i = 0] \to_p 0. \tag{A.23}$$

Consider the term $kTPR_{n,T}$ in the denominator next. Using (A.21), we have

$$kTPR_{n,T} \to_p k, \tag{A.24}$$

if $\delta > 1 - \kappa_1/3$. Using (A.23), (A.24), and noting that $(n-k)\, FPR_{n,T} \geq 0$, we have $FDR_{n,T} \to_p 0$, if $\delta > \max\{1, 2 - \kappa_1/3\}$, and $\delta^* > 2$, as required.

A.2.3 Proof of Theorem 2

We prove the error norm result first. Define a sequence $r_{\tilde{u},n}$ such that $r_{\tilde{u},n} = O(n^{3\epsilon - 3\kappa_1/2}) + O\left(n^{-\kappa_1/2}\right)$. By the definition of convergence in probability, we need to show that, for any $\varepsilon > 0$, there exists some $B_\varepsilon < \infty$, such that $\Pr\left(r_{\tilde{u},n}^{-1} |F_{\tilde{u}} - \sigma^2| > B_\varepsilon\right) < \varepsilon$. We have $\Pr\left(r_{\tilde{u},n}^{-1} |F_{\tilde{u}} - \sigma^2| > B_\varepsilon\right) \leq \Pr\left(r_{\tilde{u},n}^{-1} |F_{\tilde{u}} - \sigma^2| > B_\varepsilon | \mathcal{A}_0\right) + \Pr(\mathcal{A}_0^c)$. By (A.12), $\lim_{n\to\infty} \Pr(\mathcal{A}_0^c) = 0$. Then, it is sufficient to show that, for any $\varepsilon > 0$, there exists some $B_\varepsilon < \infty$, such that $\Pr\left(r_{\tilde{u},n}^{-1} |F_{\tilde{u}} - \sigma^2| > B_\varepsilon | \mathcal{A}_0\right) < \epsilon$. But, by (B.95) of Lemma A21 in the online supplement, the desired result follows immediately.

To prove the result for the coefficient norm, we proceed similarly. Recall that $k^* = \ominus(n^\epsilon)$ and define a sequence $r_{\beta,n}$, such that $r_{\beta,n} = O(n^{5\epsilon/2 - \kappa_1})$. To establish $\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| = O_p(r_{\beta,n})$, we need to show that, for any $\varepsilon > 0$, there exists some $B_\varepsilon < \infty$, such that $\Pr(r_{\beta,n}^{-1} \left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| > B_\varepsilon) < \varepsilon$. We have $\Pr(r_{\beta,n}^{-1} \left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| > B_\varepsilon) \leq \Pr(r_{\beta,n}^{-1} \left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| > B_\varepsilon | \mathcal{A}_0) + \Pr(\mathcal{A}_0^c)$. Again, by (A.12), $\lim_{n\to\infty} \Pr(\mathcal{A}_0^c) = 0$. Then, it is sufficient to show that, for any $\varepsilon > 0$, there exists some $B_\varepsilon < \infty$, such that $\Pr(r_{\beta,n}^{-1} \left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| > B_\varepsilon | \mathcal{A}_0 < \epsilon)$. But this follows immediately from (B.96) of Lemma A21 in the online supplement, since, conditional on the event $\mathcal{A}_0$, the set of selected regressors includes all signals.

A.2.4 Proof of Theorem 3

See Section B of the online supplement.

A.2.5 Proofs of results for the single stage OCMT in the absence of hidden signals

Result (37) follows from (25), and (38) follows from the analysis preceding Theorem 1, using (26) and (27). The result on $FDR_{n,T}$ continues to hold using the same arguments as in the proof of Theorem 1. To obtain $\Pr(\mathcal{A}_0)$ we follow the derivations in the proof of the multi-stage version of OCMT provided in Section A.2.2, but note that we only need to consider the terms from the first stage of OCMT. Similarly to (A.5) and without the need to condition on $\mathcal{D}_{k,T}$, we have $\Pr(\mathcal{A}_0^c) \leq \Pr(\sum_{i=1}^{k} \widehat{\mathcal{J}}_i < k) + \Pr(\sum_{i=k+k^*+1}^{n} \widehat{\mathcal{J}}_i > 0) = A_{n,T} + B_{n,T}$, noting that $\widehat{\mathcal{J}}_i = \widehat{\mathcal{J}}_{i,(1)}$.

Also, as with (A.9) and (A.10), we have $A_{n,T} \leq k \exp\left(-C_1 T^{C_2}\right)$. Similarly, for $B_{n,T}$ we first note that

$$B_{n,T} \leq \sum_{i=k+k^*+1}^{n} E(\widehat{\mathcal{J}}_{i,(1)} \,|\beta_i = 0\,) = \sum_{i=k+k^*+1}^{n} \Pr[|t_{\hat{\phi}_{i,(1)}}| > c_p\,(n,\delta)\,|\theta_i = 0],$$

which, by (B.51) of Lemma A10 in the online supplement, yields $B_{n,T} \leq (n-k-k^*) \exp\left[-\varkappa c_p^2(n,\delta)/2\right] + O\left[n \exp\left(-C_0 T^{C_1}\right)\right]$, or upon using result (ii) of Lemma A2, $\Pr\left(\mathcal{A}_0^c\right) \leq A_{n,T} + B_{n,T} \leq O\left(n^{1-\delta\varkappa}\right) + O\left[n \exp\left(-C_0 T^{C_1}\right)\right]$, and hence $\Pr\left(\mathcal{A}_0\right) = O\left(n^{1-\delta\varkappa}\right) + O\left[\exp\left(-n^{C_2}\right)\right]$, for some $C_2 > 0$. If, in addition, $\delta > 1$, then $\Pr\left(\mathcal{A}_0\right) \to 1$, as $n,T \to \infty$, such that $T = O\left(n^{\kappa_1}\right)$ for some $\kappa_1 > 0$, as required.

References

Antoniadis, A. and J. Fan (2001). Regularization of wavelets approximations. *Journal of the American Statistical Association 96*, 939–967.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*, 191–221.

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*, 1133–1150.

Bailey, N., M. H. Pesaran, and L. V. Smith (2016). A multiple testing approach to the regularisation of large sample correlation matrices. *CAFE Research Paper No. 14.05*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*, 29–50.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*, 608–650.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the American Statistical Association 57*, 289–300.

Berk, K. N. (1974). Consistent autoregressive spectral estimates. *Annals of Statistics 2*, 489–502.

Bickel, J. P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics 37*, 1705–1732.

Buhlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics 34*(2), 599–583.

Buhlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics 35*, 2313–2404.

Dendramis, Y., L. Giraitis, and G. Kapetanios (2015). Estimation of random coefficient time varying covariance matrices for large datasets. *Mimeo*.

Donoho, D. and M. Elad (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization. *Proceedings of the National Academy of Sciences 100*, 2197–2202.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–499.

Fan, H. L. J. and B. M. Pötscher (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics 142*, 2554–2591.

Fan, H. L. J. and B. M. Pötscher (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory 24*, 338–376.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of Royal Statistical Society B 70*, 849–911.

Fan, J. and J. Lv (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association 108*, 1044–1061.

Fan, J., R. Samworth, and Y. Wu (2009). Ultra high dimensional variable selection: Beyond the linear model. *Journal of Machine Learning Research 10*, 1829–1853.

Fan, J. and R. Song (2010). Sure independence screening in generalized linear models with np-dimensionality. *Annals of Statistics 38*, 3567–3604.

Fan, J. and C. Tang (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B 75*, 531–552.

Fithian, W., D. Sun, and J. Taylor (2014). Optimal inference after model selection. *arXiv:1410.2597v4*.

Fithian, W., J. Taylor, R. J. Tibshirani, and R. Tibshirani (2015). Selective sequential model selection. *arXiv:1512.02565*.

Freedman, D. A. (1975). On tail probabilities for martingales. *Annals of Probability 3*, 100–118.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics 29*, 1189–1232.

Friedman, J., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics 28*, 337–374.

Gavrilov, Y., Y. Benjamini, and S. K. Sarkar (2009). An adaptive step-down procedure with proven fdr control under independence. *Annals of Statistics 37*, 619–629.

G'Sell, M. G., S. Wager, A. Chouldechova, and R. Tibshirani (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B 78*, 423–444.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Huang, J., J. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics 36*, 587–613.

Li, A. and R. Barber (2015). Accumulation tests for fdr control in ordered hypothesis testing. *arXiv:1505.07352*.

Lv, J. and Y. Fan (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics 37*, 3498–3528.

Pesaran, M. H. and R. P. Smith (2014). Signs of impact effects in time series regression models. *Economics Letters 122*, 150–153.

Roussas, G. (1996). Exponential probability inequalities with some applications. *Statistica, Probability and Game Theory, IMS Lecture Notes - Monograph Series 30*, 303–319.

Stock, J. H. and M. W. Watson (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business and Economic Statistics 30*, 481–493.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani (2014). Exact post-selection inference for sequential regression procedures. *arXiv:1401.3889*.

Wecker, W. E. (1978). A note on the time series which is the product of two stationary time series. *Stochastic Processes and their Applications 8*, 153–157.

White, H. and J. M. Wooldridge (1991). Some results on sieve estimation with dependent observations. In W. J. Barnett, J. Powell, and G. Tauchen (Eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pp. 459–493. New York: Cambridge University Press.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics 38*, 894–942.

Zheng, Z., Y. Fan, and J. Lv (2014). High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society B 76*, 627–649.

Zhou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B 67*, 301–320.

Table A.1: Summary of Monte Carlo results for experiments with Gaussian innovations

	DGP-I				DGP-II				DGP-III				DGP-IV				DGP-V			
	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle*	Lasso	A-Lasso	OCMT
Static Specifications																				
TPR	1.000	0.962	0.883	0.964	1.000	0.966	0.907	0.958	1.000	0.948	0.922	0.899	1.000	0.929	0.887	0.873	1.000	0.326	0.211	0.275
FPR	0.000	0.039	0.013	0.000	0.000	0.044	0.015	0.007	0.000	0.098	0.033	0.000	0.000	0.085	0.030	0.007	0.000	0.037	0.012	0.000
FDR* (true model)	0.000	0.473	0.187	0.003	0.000	0.509	0.213	0.174	0.000	0.723	0.370	0.003	0.000	0.683	0.348	0.177	0.000	0.459	0.185	0.003
FDR (approximating model)	0.000	0.473	0.187	0.003	0.000	0.473	0.198	0.002	0.000	0.723	0.370	0.003	0.000	0.651	0.331	0.003	0.000	0.459	0.185	0.003
RMSFE	3.376	3.457	3.484	3.393	3.243	3.331	3.358	3.268	2.080	2.219	2.212	2.139	2.210	2.336	2.340	2.273	1.329	1.332	1.342	1.307
$\text{RMSE}_{\tilde{\beta}}$	0.639	0.824	1.143	0.693	0.550	0.786	1.022	0.707	0.356	0.995	0.863	0.601	0.373	0.958	0.912	0.703	0.382	0.275	0.342	0.219
Dynamic Specifications																				
Experiments with $\varphi = 0.4$																				
TPR	1.000	0.967	0.907	0.940	1.000	0.972	0.927	0.932	1.000	0.960	0.936	0.873	1.000	0.945	0.912	0.856	1.000	0.400	0.298	0.312
FPR	0.000	0.053	0.017	0.002	0.000	0.059	0.020	0.008	0.000	0.108	0.038	0.001	0.000	0.097	0.036	0.008	0.000	0.053	0.019	0.001
FDR* (true model)	0.000	0.532	0.222	0.041	0.000	0.563	0.248	0.158	0.000	0.721	0.373	0.026	0.000	0.692	0.363	0.157	0.000	0.525	0.237	0.028
FDR (approximating model)	0.000	0.518	0.217	0.013	0.000	0.518	0.229	0.002	0.000	0.705	0.365	0.003	0.000	0.647	0.340	0.003	0.000	0.512	0.232	0.003
RMSFE	3.386	3.530	3.538	3.466	3.253	3.392	3.402	3.336	2.087	2.255	2.243	2.201	2.216	2.373	2.370	2.331	1.333	1.356	1.361	1.332
$\text{RMSE}_{\tilde{\beta}}$	0.646	0.878	1.206	0.768	0.552	0.821	1.073	0.746	0.361	0.997	0.896	0.681	0.378	0.962	0.940	0.761	0.386	0.307	0.376	0.248
Experiments with $\varphi = 0.8$																				
TPR	1.000	0.962	0.881	0.868	1.000	0.968	0.896	0.847	1.000	0.946	0.887	0.810	1.000	0.935	0.874	0.786	1.000	0.394	0.282	0.250
FPR	0.000	0.049	0.016	0.028	0.000	0.054	0.018	0.021	0.000	0.101	0.031	0.017	0.000	0.090	0.027	0.020	0.000	0.050	0.013	0.017
FDR* (true model)	0.000	0.499	0.232	0.376	0.000	0.526	0.255	0.380	0.000	0.674	0.366	0.352	0.000	0.649	0.342	0.392	0.000	0.496	0.195	0.397
FDR (approximating model)	0.000	0.473	0.223	0.064	0.000	0.471	0.231	0.002	0.000	0.653	0.357	0.002	0.000	0.598	0.318	0.002	0.000	0.470	0.188	0.002
RMSFE	3.390	3.574	3.585	3.578	3.255	3.430	3.445	3.387	2.091	2.289	2.263	2.254	2.219	2.406	2.389	2.374	1.334	1.371	1.367	1.337
$\text{RMSE}_{\tilde{\beta}}$	0.645	0.877	1.130	1.313	0.551	0.819	1.010	0.993	0.360	1.030	0.900	0.929	0.376	0.985	0.906	0.971	0.383	0.303	0.366	0.294

Notes: The reported statistics represent averages across R^2 (low, medium and high), the sample sizes ($n = 100, 200, 300$ and $T = 100, 300, 500$) and all DGPs in a given design. This gives 135, 54, 27, 54 and 27 experiments for DGP-I to V, respectively. DGPs I-IV are given by (39) and DGP V is given by (40). See also notes to Table 1.

Table A.2: Summary of Monte Carlo results for experiments with non-Gaussian innovations

	DGP-I				DGP-II				DGP-III				DGP-IV				DGP-V			
	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle*	Lasso	A-Lasso	OCMT
Static Specifications																				
TPR	1.000	0.961	0.877	0.959	1.000	0.965	0.902	0.955	1.000	0.945	0.917	0.896	1.000	0.925	0.880	0.870	1.000	0.324	0.211	0.275
FPR	0.000	0.038	0.011	0.000	0.000	0.042	0.013	0.008	0.000	0.094	0.027	0.000	0.000	0.082	0.025	0.007	0.000	0.036	0.011	0.000
FDR* (true model)	0.000	0.464	0.182	0.005	0.000	0.503	0.208	0.176	0.000	0.716	0.339	0.005	0.000	0.675	0.322	0.179	0.000	0.456	0.180	0.007
FDR (approximating model)	0.000	0.464	0.182	0.005	0.000	0.467	0.192	0.004	0.000	0.716	0.339	0.005	0.000	0.643	0.304	0.005	0.000	0.456	0.180	0.007
RMSFE	3.376	3.460	3.480	3.400	3.243	3.333	3.352	3.274	2.081	2.223	2.205	2.145	2.209	2.337	2.331	2.276	1.330	1.334	1.339	1.310
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	0.648	0.833	1.138	0.716	0.558	0.798	1.017	0.733	0.362	1.007	0.853	0.624	0.379	0.970	0.904	0.723	0.388	0.281	0.332	0.232
Dynamic Specifications																				
Experiments with $\varphi = 0.4$																				
TPR	1.000	0.966	0.903	0.936	1.000	0.971	0.923	0.930	1.000	0.957	0.931	0.871	1.000	0.943	0.908	0.856	1.000	0.399	0.296	0.313
FPR	0.000	0.052	0.016	0.002	0.000	0.057	0.018	0.008	0.000	0.105	0.032	0.001	0.000	0.094	0.031	0.008	0.000	0.052	0.017	0.001
FDR* (true model)	0.000	0.528	0.215	0.042	0.000	0.557	0.239	0.159	0.000	0.714	0.341	0.026	0.000	0.686	0.338	0.159	0.000	0.521	0.228	0.030
FDR (approximating model)	0.000	0.513	0.210	0.014	0.000	0.512	0.219	0.003	0.000	0.699	0.334	0.004	0.000	0.642	0.315	0.004	0.000	0.508	0.223	0.005
RMSFE	3.386	3.532	3.532	3.473	3.255	3.395	3.397	3.342	2.089	2.260	2.234	2.207	2.217	2.376	2.362	2.334	1.334	1.356	1.356	1.334
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	0.653	0.887	1.199	0.789	0.563	0.833	1.066	0.770	0.367	1.009	0.882	0.702	0.383	0.970	0.927	0.775	0.393	0.311	0.366	0.258
Experiments with $\varphi = 0.8$																				
TPR	1.000	0.962	0.872	0.867	1.000	0.967	0.889	0.846	1.000	0.945	0.884	0.812	1.000	0.933	0.869	0.786	1.000	0.390	0.280	0.251
FPR	0.000	0.048	0.015	0.028	0.000	0.053	0.017	0.021	0.000	0.098	0.029	0.017	0.000	0.088	0.026	0.020	0.000	0.048	0.012	0.017
FDR* (true model)	0.000	0.492	0.221	0.376	0.000	0.521	0.247	0.382	0.000	0.669	0.356	0.351	0.000	0.645	0.333	0.393	0.000	0.491	0.187	0.397
FDR (approximating model)	0.000	0.466	0.213	0.064	0.000	0.465	0.223	0.002	0.000	0.647	0.347	0.002	0.000	0.594	0.308	0.002	0.000	0.465	0.181	0.003
RMSFE	3.391	3.578	3.594	3.584	3.256	3.431	3.451	3.393	2.091	2.290	2.265	2.255	2.218	2.406	2.389	2.373	1.335	1.371	1.366	1.339
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	0.654	0.885	1.146	1.328	0.560	0.832	1.031	1.007	0.366	1.040	0.916	0.933	0.383	0.996	0.922	0.981	0.391	0.308	0.366	0.302

Notes: See notes to Tables 1 and A.1.